Connecting Ansatz Expressibility to Gradient Magnitudes and Barren Plateaus

Zoë Holmes,[1,*] Kunal Sharma,[2,3] M. Cerezo,[3,4] and Patrick J. Coles[3]

[1] *Information Sciences, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

[2] *Hearne Institute for Theoretical Physics, Department of Physics and Astronomy, and Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803, USA*

[3] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

[4] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

(Received 12 February 2021; accepted 30 November 2021; published 24 January 2022)

Parametrized quantum circuits serve as ansatze for solving variational problems and provide a flexible paradigm for the programming of near-term quantum computers. Ideally, such ansatze should be highly expressive, so that a close approximation of the desired solution can be accessed. On the other hand, the ansatz must also have sufficiently large gradients to allow for training. Here, we derive a fundamental relationship between these two essential properties: expressibility and trainability. This is done by extending the well-established barren plateau phenomenon, which holds for ansatze that form exact 2-designs, to arbitrary ansatze. Specifically, we calculate the variance in the cost gradient in terms of the expressibility of the ansatz, as measured by its distance from being a 2-design. Our resulting bounds indicate that highly expressive ansatze exhibit flatter cost landscapes and therefore will be harder to train. Furthermore, we provide numerics illustrating the effect of expressibility on gradient scalings and we discuss the implications for designing strategies to avoid barren plateaus.

DOI: 10.1103/PRXQuantum.3.010313

I. INTRODUCTION

While quantum hardware is rapidly reaching the stage at which it can outperform classical supercomputers [1], we remain in the noisy intermediate-scale quantum (NISQ) era, in which the available devices are relatively small and prone to errors [2]. Variational quantum algorithms (VQAs) have gathered attention as a computational strategy that is well suited to the constraints imposed by NISQ devices [3–20]. In VQAs, a problem-specific cost function is efficiently evaluated on a quantum computer, while a classical optimizer trains a parametrized quantum circuit to minimize this cost. The benefit of this paradigm is that it adapts to the qubit and connectivity constraints of NISQ devices, while keeping the circuit depth short to mitigate quantum hardware noise.

Central to the success of VQAs is the construction of a parametrized quantum circuit, which serves as an ansatz with which to explore the space of solutions to the target problem. Some noteworthy ansatze include the quantum alternating operator ansatz [5,21], the coupled cluster ansatz [22–24], the Hamiltonian variational ansatz [25], and the hardware-efficient ansatz [26]. To successfully find an optimal solution, the ansatz should ideally be both expressive and trainable. Specifically, the ansatz must be sufficiently expressive that it contains a circuit that approximates the optimal solution well. Concurrently, the cost landscape must be sufficiently featured to be able to train the parameters to find this optimal solution.

Recently, it has been shown that VQAs can exhibit barren plateaus, where under certain conditions the gradient of the cost function vanishes exponentially with the size of the system [27–36]. In particular, Ref. [27] has demonstrated that if an ansatz is sufficiently random that it matches the uniform distribution of unitaries up to the second moment (i.e., forms a 2-design), then the variance in the cost gradient will vanish exponentially with the number of qubits. Several strategies have been proposed to address this issue [37–46], such as clever parameter initialization or ansatz construction, while more research is needed to test these strategies on various problems.

In broad terms, the expressibility of an ansatz is determined by how uniformly it explores the unitary space. Thus the distance between the distribution of unitaries generated by an ansatz and the maximally expressive uniform distribution of unitaries is a natural measure of its expressibility [47]. Using such a measure, Ref. [48] has

[*] zholmes@lanl.gov

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

calculated the expressibility for several commonly used ansatze and, by using the cost gradients obtained in Ref. [28], has suggested that in some cases it is possible for an ansatz to be both expressive and trainable. Additionally, Ref. [49] has noted a numerical correlation between expressibility and trainability for analog systems. However, given that both expressibility and trainability are closely related to randomness, one might expect to be able to draw a more fundamental and general relationship between expressibility and trainability.

Here, we demonstrate that this is indeed the case by analytically relating the trainability of an ansatz to its expressibility. This is done by extending the barren plateau phenomenon introduced in Ref. [27], which holds for ansatze that form exact 2-designs, to arbitrary ansatze. Specifically, we upper bound the variance in the cost gradient in terms of the distance the ansatz is from being a 2-design. Since the degree to which an ansatz is a 2-design is a measure of its expressibility, this allows us to relate the gradient of the cost landscape to the expressibility of the ansatz. We find that the more expressive the ansatz, the smaller is the variance in the cost gradient and hence the flatter is the landscape. We note that an ansatz does not strictly need to be highly expressive to be used successfully; rather, it just needs to contain a solution to the problem at hand. Thus our result highlights the importance of developing trainable problem-inspired ansatze.

Our main results can be summarized in Fig. 1. Given an ansatz, we analyze the space of unitaries that are accessible when sampling the parameters [Fig. 1(a)] of a parametrized quantum circuit. Inexpressive ansatze, such as the one shown in Fig. 1(b), access a small region of the unitary group and can include the space of unitaries that solve certain problems but not the space that solve others. Our results do not preclude inexpressive ansatze having trainability issues, such as barren plateaus. On the other hand, highly expressive ansatze, which are generically used for many problems, as they can access a much larger space [Fig. 1(c)], are shown to lead to small gradients and hence can have trainability issues.

Since our analytic bounds are upper bounds, they leave open the questions of how reducing the expressibility of an ansatz changes the cost landscape and hence how reducing the expressibility can be used to avoid the barren plateau phenomenon. To address these questions, we provide extensive numerics studying the effect that tuning the expressibility of an ansatz may have on the scaling of gradient magnitudes. Specifically, we consider the effects of decreasing the depth of the circuits, correlating circuit parameters, and restricting either the direction or the angle of rotations. We find that strongly correlating parameters [37] and/or initializing close to the solution (and then restricting the ansatz to explore the region close to the initialization [38]) are the most effective approaches to avoid exponentially vanishing cost gradients.

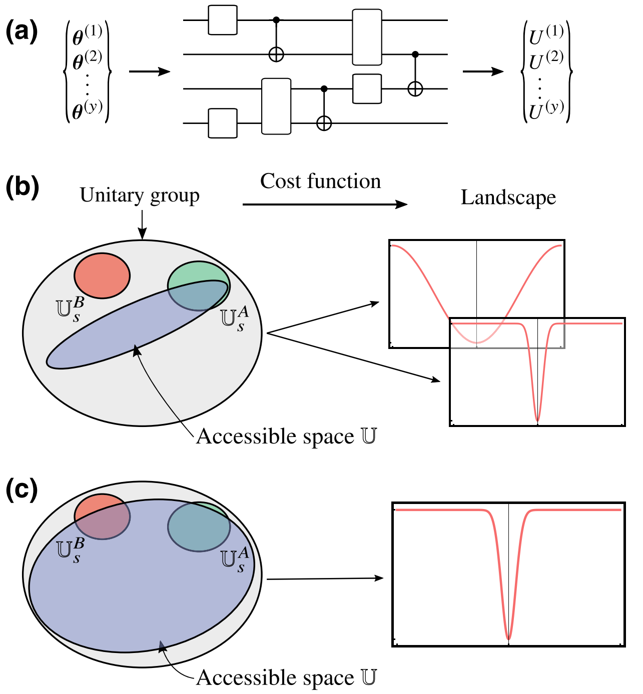

FIG. 1. A schematic representation of the main results. (a) Variational quantum algorithms (VQAs) train the parameters $\boldsymbol{\theta}$ in a parametrized quantum circuit to minimize a cost function as in Eq. (1). Each set of parameters corresponds to a unitary $U(\boldsymbol{\theta})$ being produced. The set of unitaries $\mathbb{U}$ accessible by $U(\boldsymbol{\theta})$ is a subset of the unitary group $\mathcal{U}(d)$ and the VQA can be successful if $\mathbb{U}$ overlaps with the space of solution unitaries $\mathbb{U}_s$ that (approximately) minimize the cost. The expressibility of an ansatz quantifies the degree to which it uniformly explores the unitary group $\mathcal{U}(d)$. Given problems A and B, we denote their solution spaces as $\mathbb{U}_s^A$ and $\mathbb{U}_s^B$, respectively. (b) A low-expressibility ansatz contains solutions to problem A but not to B, while a high-expressibility ansatz as in (c) contains solutions to both problems. Low-expressibility ansatze can lead to both small and large cost gradients. On the other hand, high-expressibility ansatze lead to predominantly flat cost landscapes and thus are generally hard to train.

II. PRELIMINARIES

A. General framework

VQAs encode an optimization task in a cost function, the minimum of which corresponds to the solution of the problem. Here, we consider cost functions of the form [50]

$$C_{\rho,H}(\boldsymbol{\theta}) = \mathrm{Tr}[HU(\boldsymbol{\theta})\rho U(\boldsymbol{\theta})^\dagger], \qquad (1)$$

where ρ is an n-qubit input state, H is a Hermitian operator, and $U(\boldsymbol{\theta})$ is a parametrized quantum circuit depending on trainable parameters $\boldsymbol{\theta}$. The value of the cost $C_{\rho,H}(\boldsymbol{\theta})$ (or of its gradient) are estimated on a quantum computer and are then fed into a classical optimizer, which attempts to solve the optimization task arg $\min_{\boldsymbol{\theta}} C_{\rho,H}(\boldsymbol{\theta})$.

The success of the VQA hinges on several factors. First, it is necessary to find an operator H such that the resulting cost is faithful for the given problem. That is, we require the minimum of $C_{\rho,H}(\boldsymbol{\theta})$ to correspond to the solution of the optimization task. Evidently, for some applications, there may be multiple choices in H corresponding to faithful costs and therefore other factors will determine which to use. One such factor is how easily H can be measured on a quantum computer. Another relevant feature, as discussed further in Sec. II D, is the *locality* of H, i.e., the number of qubits on which it acts nontrivially. We say that the cost function is *global* if H acts nontrivially on all qubits, while we use the term *k-local* for costs where H acts nontrivially on at most k qubits.

A second aspect that determines the success of a VQA is the choice of *ansatz* for $U(\boldsymbol{\theta})$. While discrete parametrizations are possible, usually $\boldsymbol{\theta}$ are continuous parameters, such as gate rotation angles, in a parametrized quantum circuit. Generally, $U(\boldsymbol{\theta})$ is expressed as

$$U(\boldsymbol{\theta}) = \prod_{j=1}^{D} U_j(\theta_j) W_j . \tag{2}$$

Here, $\{W_j\}_{j=1}^N$ is a chosen set of fixed unitaries and $U_j = e^{-i\theta_j V_j}$ is a rotation of angle θ_j generated by a Hermitian operator V_j such that $(V_j)^2 = \mathbb{1}$. The rotation angles $\{\theta_j\}$ are typically assumed to be independent.

Once an ansatz has been fixed for the parametrized quantum circuit, then, as sketched in Fig. 1(a), each possible vector of parameters $\boldsymbol{\theta}$ corresponds to a unitary $U(\boldsymbol{\theta})$ that is produced. For concreteness, given a set of different parameters $\{\boldsymbol{\theta}^{(1)}, \ldots \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(y)}\}$ we obtain the corresponding ensemble of unitaries:

$$\mathbb{U} = \{U^{(1)}, U^{(2)}, \ldots, U^{(y)}\}, \tag{3}$$

where $U^{(j)} := U(\boldsymbol{\theta}^{(j)})$. Here, $\mathbb{U} \subseteq \mathcal{U}(d)$, where $\mathcal{U}(d)$ is the unitary group $\mathcal{U}(d)$ of degree $d = 2^n$.

B. Expressibility

For a VQA to be successful, a solution (i.e., a unitary that is by some measure close to the unitary that minimizes the cost) needs to be contained within the ensemble of unitaries generated by the ansatz. Specifically, defining $\mathbb{U}_s$ as the set of solution unitaries, then the VQA will be successful only if $\mathbb{U}_s \bigcap \mathbb{U} \neq \emptyset$. When this condition is satisfied, the ansatz is said to be *complete* for the given problem.

In the absence of prior knowledge about where the solution unitaries $\mathbb{U}_s$ lie, the likelihood that the ansatz is complete can be maximized by using an ansatz that explores the total space of unitaries as fully and as uniformally as possible. Such ansatze are known as *expressive* ansatze. For example, consider having two problems (problems A and B), with solution spaces respectively denoted as $\mathbb{U}_s^A$ and $\mathbb{U}_s^B$. Figure 1(b) sketches $\mathbb{U}$ for an inexpressive ansatz that is complete with respect to problem A but incomplete with respect to B. Conversely, Fig. 1(c) shows $\mathbb{U}$ for an expressive ansatz that is complete with respect to both problems.

For many applications, information about the problem can be encoded in the ansatz. For instance, the quantum alternating operator ansatz [21] (or the Hamiltonian variational ansatz [25]), encodes information of an appropriate adiabatic transformation. Such *problem-inspired* ansatze may be complete but inexpressive [e.g., Fig. 1(b) could denote a problem-inspired ansatz for problem A]. However, *problem-agnostic* ansatze, which can be used for a wide range of problems, need to be sufficiently expressive to guarantee their completeness.

The expressibility of an ansatz, i.e., the degree to which it uniformly explores the unitary group $\mathcal{U}(d)$, can be quantified by comparing the uniform distribution of unitaries obtained from the ensemble $\mathbb{U}$ to the maximally expressive uniform (Haar) distribution of unitaries from $\mathcal{U}(d)$. More concretely, the expressibility of a circuit can be defined in terms of the following superoperator [47,48]:

$$\begin{aligned}\mathcal{A}_{\mathbb{U}}^{(t)}(\cdot) := &\int_{\mathcal{U}(d)} d\mu(V) V^{\otimes t}(\,\cdot\,)(V^{\dagger})^{\otimes t} \\ &- \int_{\mathbb{U}} dU\, U^{\otimes t}(\,\cdot\,)(U^{\dagger})^{\otimes t},\end{aligned} \tag{4}$$

where $d\mu(V)$ is the volume element of the Haar measure and dU is the volume element corresponding to the uniform distribution over $\mathbb{U}$ in Eq. (3). If $\mathcal{A}_{\mathbb{U}}^{(t)}(X) = 0$ for all operators X, then averaging over elements of $\mathbb{U}$ agrees with averaging over elements of the Haar distribution over $\mathcal{U}(d)$ up to the tth moment, and thus $\mathbb{U}$ forms a t-design [51–55]. For our purposes, it suffices to consider the behavior of $\mathcal{A}_{\mathbb{U}}^{(t)}$ for $t = 2$. Henceforth, we drop the t superscript; i.e., $\mathcal{A}_{\mathbb{U}} \equiv \mathcal{A}_{\mathbb{U}}^{(2)}$.

In the context of minimizing a generic cost $C_{\rho,H}(\boldsymbol{\theta})$ of the form specified by Eq. (1), we are interested in the expressibility of the circuit with respect to both the initial state ρ and the measurement operator H. The following quantities, respectively, capture these notions:

$$\varepsilon_{\mathbb{U}}^{\rho} := ||\mathcal{A}_{\mathbb{U}}(\rho^{\otimes 2})||_2, \tag{5}$$

$$\varepsilon_{\mathbb{U}}^{H} := ||\mathcal{A}_{\mathbb{U}}(H^{\otimes 2})||_2. \tag{6}$$

Small values of $\varepsilon_{\mathbb{U}}^{\rho}$ and $\varepsilon_{\mathbb{U}}^{H}$ indicate that the ansatz is highly expressive. These measures generalize the notion of expressibility introduced in Ref. [47], where the expressibility has been defined in terms of $\varepsilon_{\mathbb{U}}^{\rho}$ for $\rho = |0\rangle\langle 0|$.

While the ρ and H dependence of $\varepsilon^{\rho}_{\mathbb{U}}$ and $\varepsilon^{H}_{\mathbb{U}}$ make them natural measures of the expressibility in the context of minimizing a cost $C_{\rho,H}(\boldsymbol{\theta})$, cost-function-independent measures of expressibility may allow the expected performance of different ansatze to be more easily compared. With this in mind, one could alternatively quantify the expressibility directly in terms of the diamond norm of $\mathcal{A}_{\mathbb{U}}$,

$$\varepsilon^{\diamond}_{\mathbb{U}} := ||\mathcal{A}_{\mathbb{U}}||_{\diamond}, \tag{7}$$

which is an operationally meaningful distance measure to distinguish two quantum operations. We use the diamond norm here in line with the literature on ε-approximate unitary designs [56]; however, alternative norms can be used (for a discussion, see Ref. [54]). For completeness, we formulate our results in terms of $\varepsilon^{\diamond}_{\mathbb{U}}$, as well as the quantities $\varepsilon^{\rho}_{\mathbb{U}}$ and $\varepsilon^{H}_{\mathbb{U}}$.

C. Gradient magnitudes

For a VQA to run successfully, it is not sufficient that the ansatz contains the solution; the cost landscape must also exhibit large enough cost gradients to enable this solution to be found.

The component of the gradient corresponding to the parameter θ_k is determined by the partial derivative $\partial_k C := \partial C_{\rho,H}(\boldsymbol{\theta})/\partial\theta_k$. For a generic ansatz of the form specified by Eq. (2), the average of $\partial_k C$ over all parameters $\boldsymbol{\theta}$ vanishes:

$$\langle \partial_k C \rangle = 0 \quad \forall \;\; k. \tag{8}$$

That is, the cost gradients are not biased in any single direction but, rather, average out to zero. Intuitively, this lack of bias can be understood as following from the fact that the average of a rotation $\exp[-i\theta_k V_k]$ is zero when $V_k^2 = \mathbb{1}$. We show this in Appendix C, where we prove that $\langle \partial_k C \rangle = 0$ by explicitly integrating over θ_k.

However, an unbiased cost landscape can be either trainable or untrainable, depending on the extent to which the gradient fluctuates away from zero. Therefore, to assess the trainability of an ansatz $U(\boldsymbol{\theta})$, we now recall Chebyshev's inequality. This inequality bounds the probability that the partial derivative of the cost deviates from its average of zero,

$$P(|\partial_k C| \geqslant \delta) \leqslant \frac{\mathrm{Var}[\partial_k C]}{\delta^2}, \tag{9}$$

in terms of the variance

$$\mathrm{Var}[\partial_k C] = \left\langle (\partial_k C)^2 \right\rangle - \langle \partial_k C \rangle^2, \tag{10}$$

where the expectation value is taken over the parameters $\boldsymbol{\theta}$. Hence if the variance of the partial derivative is small for all θ_k, then the probability that the partial derivative is nonzero is small for all θ_k. On such landscapes, (potentially untenably) precise measurements are required to detect the path of steepest descent to navigate to the minimum.

D. Barren plateaus

There is a growing awareness of the so-called barren plateau phenomenon for VQAs [27–36]. For a given ansatz $U(\boldsymbol{\theta})$, a cost C is said to exhibit a barren plateau if its gradients vanish exponentially with the number of qubits n. This is typically relaxed to a probabilistic definition, where the gradient vanishes exponentially with high probability. This would follow from Chebyshev's inequality, given in Eq. (9), if the variance in the partial derivative vanishes exponentially, i.e., if $\mathrm{Var}[\partial_k C] \in \mathcal{O}(2^{-pn})$ for any integer $p > 0$. For costs that exhibit barren plateaus, exponentially precise measurements may be required to determine the minimization direction and hence the cost is effectively untrainable for large problem sizes.

To elucidate the conditions under which a layered parametrized ansatz $U(\boldsymbol{\theta})$, of the form of Eq. (2), gives rise to barren plateaus, consider a bipartite cut of $U(\boldsymbol{\theta})$ and write

$$U(\boldsymbol{\theta}) = U_L(\boldsymbol{\theta})U_R(\boldsymbol{\theta}), \tag{11}$$

where

$$U_L(\boldsymbol{\theta}) = \prod_{j=k+1}^{D} U_j(\boldsymbol{\theta}_{\mathbf{j}})W_j \quad \text{and} \quad U_R(\boldsymbol{\theta}) = \prod_{j=1}^{k} U_j(\boldsymbol{\theta}_j)W_j. \tag{12}$$

Note that since we suppose that the parameters θ_j are uncorrelated, the circuits U_L and U_R are independent. These circuits are pertinent when quantifying gradients, since taking the partial derivative of a circuit, as shown in Appendix D, effectively splits a circuit in two.

Reference [27] has then demonstrated that if the ensemble of unitaries generated by the ansatz $U(\boldsymbol{\theta})$ is sufficiently random (i.e., expressive) such that the ensembles $\mathbb{U}_L$ or $\mathbb{U}_R$ [associated with the circuits $U_L(\boldsymbol{\theta})$ and $U_R(\boldsymbol{\theta})$, respectively] form 2-designs, then the variance in the cost gradient vanishes exponentially with n. Specifically, let us denote the variance of the cost when just $\mathbb{U}_R$, just $\mathbb{U}_L$, and both $\mathbb{U}_R$ and $\mathbb{U}_L$ form 2-designs as $\mathrm{Var}_R\partial_k C$, $\mathrm{Var}_L\partial_k C$, and $\mathrm{Var}_{R,L}\partial_k C$, respectively. From Ref. [27], it follows that for $x = R$, $x = L$ and $x = R, L$,

$$\mathrm{Var}_x\partial_k C = \frac{g_x(\rho, H, U)}{2^{2n} - 1}, \tag{13}$$

where we pull out the n-dependent scaling factor explicitly. The prefactor $g_x(\rho, H, U)$, which we define explicitly in Appendix E, is in $\mathcal{O}(2^n)$ for typical choices in V_k and H. Thus if $\mathbb{U}_L$ or $\mathbb{U}_R$ form a 2-design, the variance in the gradient vanishes exponentially in n. In other words, maximally expressive ansatze exhibit barren plateaus.

III. MAIN RESULTS

A. Analytic bounds

In this section, we study the gradient of a generic cost $C_{\rho,H}(\boldsymbol{\theta})$, given in Eq. (1), with an ansatz $U(\boldsymbol{\theta})$, given in Eq. (2), but relax the assumption that $\mathbb{U}_L$ or $\mathbb{U}_R$ forms a 2-design. By doing so, we extend the results on barren plateaus from Ref. [27] to arbitrary ansatze. As will become clear, this generalization enables us to relate the variance of the cost function partial derivative to the expressibility of $U(\boldsymbol{\theta})$ in Eq. (4).

Let us start by noting that while maximally expressive ansatze exhibit barren plateaus, the converse is not necessarily true. In other words, highly inexpressive ansatze need not always experience large cost gradients and in fact they may exhibit vanishing gradients. A trivial example of this phenomenon is provided by an ansatz composed of rotations that commute with the measurement operator $[U(\theta), H] = 0$. Such an ansatz will leave the cost unchanged for any θ and so the variance in gradient in the cost of such an ansatz is necessarily zero. A more subtle example is an ansatz composed of a tensor product of single-qubit rotations. Since this ansatz does not generate entanglement, it is inexpressive; however, it has also been shown to exhibit a barren plateau for global cost functions [7,28]. It follows from these observations that it is not possible to meaningfully *lower* bound the gradients of an ansatz in terms of its expressibility.

Therefore, to relate cost gradients to expressibility, we instead derive an *upper* bound. Specifically, our main result consists of a nontrivial upper bound for the variance of the cost function partial derivative for a general ansatz $U(\boldsymbol{\theta})$ in terms of the expressibility in Eq. (4). This bound is in terms of (1) the variance of the cost gradient when either $\mathbb{U}_L$ or $\mathbb{U}_R$ form a 2-design, and (2) the expressibility of the ansatz as measured by the distance $\mathbb{U}_L$ and $\mathbb{U}_R$ are from being 2-designs. As shown in Appendix D, we prove the following.

Theorem 1. *Consider a generic cost function $C_{\rho,H}(\boldsymbol{\theta})$, as given in Eq. (1), using a layered ansatz $U(\boldsymbol{\theta})$ of the general form in Eq. (2). The variance of the cost partial derivative obeys the following bounds:*

$$Var\ \partial_k C \leqslant Var_R\ \partial_k C + 4\varepsilon_R^{\rho}||H||_2^2, \tag{14}$$

$$Var\ \partial_k C \leqslant Var_L\ \partial_k C + 4\varepsilon_L^{H}||\rho||_2^2, \tag{15}$$

$$Var\ \partial_k C \leqslant Var_{R,L}\ \partial_k C + f\,(\varepsilon_R^{\rho}, \varepsilon_L^{H}). \tag{16}$$

Here, we use the shorthand $\varepsilon_R^{\rho} := \varepsilon_{\mathbb{U}_R}^{\rho}$ and $\varepsilon_L^{H} := \varepsilon_{\mathbb{U}_L}^{H}$, and we define

$$f\,(x,y) := 4xy + \frac{2^{n+2}\left(x||H||_2^2 + y||\rho||_2^2\right)}{2^{2n}-1}. \tag{17}$$

Theorem 1 establishes a formal relationship between the gradient of the cost landscape and the expressibility of the ansatz used. Namely, the higher the expressibility of the ansatz—that is, the smaller ε_L^H or ε_R^{ρ}—the smaller is the upper bound on the variance of the cost partial derivative. This, in combination with the fact that the cost gradient is unbiased, demonstrates that highly expressive ansatze will have flatter landscapes and consequently be harder to train.

In contrast to the bounds specified by Eq. (13), which hold for three distinct cases (i.e., when $\mathbb{U}_L$ is a 2-design, when $\mathbb{U}_R$ is a 2-design, and when both $\mathbb{U}_L$ and $\mathbb{U}_R$ are 2-design), the bounds in Eqs. (14)–(16) all hold for any generic ansatz of the form in Eq. (2). Thus any single bound would suffice to bound the variance in the cost function partial derivative for an arbitrary ansatz.

We include all three bounds despite this fact since, in any instance, one bound may be tighter than the others and hence more informative. In particular, the relative tightness of the bounds depends on the parameter with respect to which we are taking the derivative. This follows from the fact that Eq. (14) becomes an equality in the limit that $\mathbb{U}_R$ tends to a 2-design, whereas Eq. (15) becomes an equality in the limit that $\mathbb{U}_L$ is a 2-design and Eq. (16) becomes an equality in the limit that both $\mathbb{U}_L$ and $\mathbb{U}_R$ are 2-designs. If we are looking at the derivative with respect to the final layer, then $\mathbb{U}_R$ is typically closer to being a 2-design than $\mathbb{U}_L$ and so Eq. (14) will be tightest. Conversely, if we are most interested in the partial derivative with respect to a parameter in the first layer, then Eq. (15) will be tightest. On the other hand, for parameters in a layer close to the middle (i.e., at depth $D/2$) and Eq. (16) will be tightest since, as shown in Appendix D, the derivation of this bound uses the most information about the ansatz.

In Appendix D, we extend Theorem 1 to cost functions of the form $C_{\text{gen}} = \sum_i \text{Tr}[H_i U(\boldsymbol{\theta})\rho_i U(\boldsymbol{\theta})^{\dagger}]$, which allow for multiple input states and measurements. Thus our results also apply to quantum machine learning approaches that utilize training data [57–60].

1. Generalizing the barren plateau phenomenon

Theorem 1 may be viewed as an extension of the barren plateau phenomenon introduced in Ref. [27] to ansatze that form approximate rather than exact 2-designs. By combining Eqs. (13) and (16), we find that the variance in the partial derivative for an arbitrary ansatz is bounded as

$$\text{Var}\ \partial_k C \leqslant \frac{g_{L,R}(\rho, H, U)}{2^{2n}-1} + f\,(\varepsilon_L^{H}, \varepsilon_R^{\rho}). \tag{18}$$

Here, the first term on the right is the variance of a maximally expressive ansatz (namely, one that forms a 2-design) and $f\,(\varepsilon_L^H, \varepsilon_R^{\rho})$ is the expressibility-dependent correction term defined in Eq. (17). Expressions similar to Eq. (18) are obtainable from Eqs. (14) and (15).

For perfectly expressive ansatze, $f(\varepsilon_L^H, \varepsilon_R^\rho)$ vanishes and Eq. (18) reduces to Eq. (13), regaining the result of Ref. [27]. In this case, the variance in the gradient vanishes exponentially with the size of the system n, i.e., the ansatz exhibits a barren plateau. Similarly, if the expressibility of an ansatz increases exponentially with the size of the problem, i.e., if $f(\varepsilon_L^H, \varepsilon_R^\rho) \in \mathcal{O}\left(1/2^{kn}\right)$ for $k > 0$, then Var $\partial_k C$ again vanishes exponentially and the ansatz exhibits a barren plateau. However, more generally, when $f(\varepsilon_L^H, \varepsilon_R^\rho)$ scales nonexponentially, the upper bound allows for the variance in the partial derivative to be nonvanishing. Thus, there is leeway for imperfectly expressive ansatze to avoid barren plateaus.

In Ref. [61], it has been proven that the barren plateau phenomenon is necessarily associated with the concentration of cost function values about their mean. More concretely, it has been shown that the probability that the cost function deviates from its mean is determined by the variation in the gradient of the cost. Thus our bounds also imply that the degree to which the cost concentrates about its mean increases with increasing expressibility. In Appendix F, we provide an alternative proof of this following on from the results of Ref. [33].

2. Diamond norm reformulation

For local costs, the term $||H||_2^2$ scales exponentially with the size of the system and therefore for large systems Eq. (14) becomes exponentially loose. This issue can be mitigated by reformulating Theorem 1 in terms of $\varepsilon_{\mathbb{U}}^\diamond$, given in Eq. (7). We obtain the following theorem in Appendix D.

Theorem 2. *Consider a generic cost function $C_{\rho,H}(\boldsymbol{\theta})$, as given in Eq. (1), using a layered ansatz $U(\boldsymbol{\theta})$ of the general form in Eq. (2). The variance of the cost partial derivative obeys the following bounds:*

$$\mathit{Var}\ \partial_k C \leqslant \mathit{Var}_R\ \partial_k C + 4||H||_\infty^2\ \varepsilon_R^\diamond, \tag{19}$$

$$\mathit{Var}\ \partial_k C \leqslant \mathit{Var}_L\ \partial_k C + 4||\rho||_\infty^2 \|H\|_1\ \varepsilon_L^\diamond, \tag{20}$$

$$\mathit{Var}\ \partial_k C \leqslant \mathit{Var}_{R,L}\ \partial_k C + \frac{f(\varepsilon_R^\diamond, \|H\|_1 \varepsilon_L^\diamond)}{2^n}, \tag{21}$$

where we use the shorthand $\varepsilon_R^\diamond = \varepsilon_{\mathbb{U}_R}^\diamond$ and $\varepsilon_L^\diamond = \varepsilon_{\mathbb{U}_L}^\diamond$ and with $f(x,y)$ defined in Eq. (17).

Again, Theorem 2 formally establishes that highly expressive ansatze experience flatter cost landscapes. Furthermore, a relation similar to Eq. (18) can be derived from Theorem 2. Hence, Theorem 2 also provides an extension of the barren plateau result of Ref. [27]. However, since $||H||_\infty^2 \in \mathcal{O}(1)$ for all H, Eq. (19) does not experience the same looseness for local costs of large systems as Eq. (14). On the other hand, since $||H||_1$ may scale exponentially in n, Eq. (20) may become loose for large systems and therefore we expect Eq. (15) to generally be more useful than Eq. (20).

B. Numerical simulations

Since the analytic bounds in the previous section are upper bounds, we have no guarantee that inexpressive ansatze will exhibit larger cost gradients. The bounds thus leave open the question of whether or how reducing the expressibility of an ansatz changes the cost landscape. Moreover, they leave open the question of how one can avoid the barren plateau phenomenon that is observed for maximally expressive ansatze.

One can conceive of numerous ways in which the expressibility of an ansatz can be tuned, each of which could have a different impact. In this section, we consider four such ways: decreasing the depth of the circuits, correlating circuit parameters, and restricting either the direction or angle of rotations. We then numerically investigate the effect these have on the cost gradient scaling.

For completeness, in our numerics we consider both a 2-local cost where the measurement operator is composed of Pauli-z measurements on the first and second qubits, $H_L = \sigma_1^z \sigma_2^z$, and a global cost where the measurement operator consists of Pauli-z measurements across all qubits, $H_G = \bigotimes_{i=1}^n \sigma_i^z$ [28]. In both cases, following Ref. [27], the system is prepared in the pure state, $\rho = |\psi_0\rangle\langle\psi_0|^{\otimes n}$, where $|\psi_0\rangle = \exp[-i(\pi/8)\sigma_Y]|0\rangle$. We further consider a layered hardware-efficient ansatz,

$$U(\mathbf{k}_l, \boldsymbol{\theta}_l, D) := \prod_{l=1}^{D} WV(\mathbf{k}_l, \boldsymbol{\theta}_l), \tag{22}$$

consisting of D alternating layers of random single-qubit gates and entangling gates as shown in Fig. 2. Specifically, the entangling layer,

$$W = \prod_{i=1}^{n-1} \text{CPHASE}_{i,i+1}, \tag{23}$$

is composed of a ladder of controlled-phase (CPHASE) operations, between adjacent qubits in a one-dimensional array. The single-qubit layer consists of a series of random single-qubit rotations,

$$V(\mathbf{k}_l, \boldsymbol{\theta}_l) = \prod_{i=1}^{n} R_{k_l^i}(\theta_l^i), \tag{24}$$

where $R_{k_l^i}(\theta_l^i)$ is a rotation of the i_{th} qubit by an angle θ_l^i about the $k_l^i = x, y$ or z axis. In the maximally expressive version of the ansatz, the x, y, or z rotation directions $\{k_l^i\}$ for each qubit on each layer are chosen independently

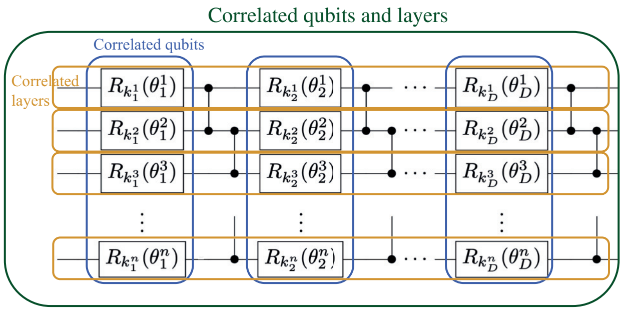

FIG. 2. Ansatz employed in numerical simulations. The ansatz is composed of alternating random single-qubit rotations and ladders of CPHASE operations. The colored boxes indicate the gates, which are fixed to rotate by the same angle and in the same direction when we correlate the ansatz layers (yellow), correlate qubits (blue), and correlate both the layers and qubits (green).

and with equal probability and the rotation angles $\{\theta_l^i\}$ are independently and randomly chosen in the range 0 to 2π. Our numerics are implemented using TensorFlow Quantum [62].

1. Circuit depth

One of the simplest ways of reducing the expressibility of an ansatz is to reduce the depth D of the circuit. It has been shown in Ref. [28] that global costs with a hardware-efficient ansatz experience barren plateaus irrespective of the depth of the circuit. However, local costs only exhibit barren plateaus for deep circuits $D \in \Omega(\text{poly}(n))$ but are trainable for shallow circuits $D \in \mathcal{O}(\log(n))$.

We obtain similar results here. As shown in Fig. 3(a), for the global cost, the variance in the partial derivative is seemingly independent of the depth of the circuit and vanishes exponentially with the size of the system n. Conversely, for local costs, as shown in Fig. 3(d), exponentially vanishing partial derivatives are observed for systems up to 12 qubits for depths $D \gtrsim 100$. However, shallow circuits $D \lesssim 50$ exhibit an approximately constant scaling for $n \gtrsim 8$.

2. Correlating parameters

A more sophisticated means of reducing the expressibility of the ansatz is to correlate the rotation angles [37]. Here, we consider three different means of correlating parameters, as sketched in Fig. 2, and plot the corresponding variance in the cost partial derivative in the central panel of Fig. 3. In the first, shown in yellow, we correlate the qubits (but allow the angles to vary between layers), i.e., $k_l^i = k_l^{i'}$ and $\theta_l^i = \theta_l^{i'}$ for any two qubits i and i'. In the second (plotted in green), we correlate the different layers (but not the qubits), i.e., $k_l^i = k_{l'}^i$ and $\theta_l^i = \theta_{l'}^i$ for any two layers l and l'. Finally, as shown in blue, we correlate both the qubits and layers. In this case, all the qubits rotate in same direction and by the same angle, i.e., $k_l^i = k_{l'}^{i'}$ and $\theta_l^i = \theta_{l'}^{i'}$ for any two qubits i and i' and layers l and l'. In other words, all parameters are correlated. The data for only y (x) rotations are indicated by the solid (dashed) lines, respectively.

In contrast to varying circuit depth, here we obtain similar results irrespective of whether a local or global cost is used. Correlating both the qubits and the layers results in the least expressive ansatz and, correspondingly, the largest variation in cost gradients is observed. Indeed, in this case the variance in the cost gradient is approximately constant. In contrast, correlating just the qubits, or just the layers, increases the cost gradients and reduces the scaling of the cost gradient with system size but an exponential scaling is still observed.

3. Restricting rotation direction

One might also consider reducing the expressibility of the ansatz by reducing the single-qubit-rotation gates to a subset of directions. We explore this in the right panel of Fig. 3. In blue, we plot the variance when only rotations in a single direction, namely in the x (dark blue) or y (light blue) direction, are implemented. We do not plot the case when only z rotations are implemented, since in that case U commutes with $H_L = \sigma_1^z \sigma_2^z$ and $H_G = \bigotimes_{i=1}^n \sigma_i^z$, and so the cost landscape is entirely flat. For a local cost, reducing the expressibility of the ansatz by restricting to single-direction rotations seemingly removes the exponential gradient scaling. However, for a global cost, the scaling remains exponential.

4. Restricting rotation angles

A final way to reduce the expressibility of an ansatz is by reducing the range from which the rotation angles $\boldsymbol{\theta}$ are chosen; that is, choosing the θ_l^i in the range $[\tilde{\theta}_l^i, \tilde{\theta}_l^i + 2\pi r]$, where $\tilde{\theta}_l^i$ is a fixed initialization point. For $r = 1$, the ansatz explores the entire solution space but for $r < 1$ the ansatz is constrained to exploring a subset of the solution space where the rotation angles θ_l^i deviate from $\tilde{\theta}_l^i$ by at most $2\pi r$.

However, with a little thought, it is clear that, in contrast to the previous three approaches we have discussed, restricting the rotation angles of the ansatz does not change the cost landscape but, rather, limits the region of the landscape explored by the ansatz. Thus, in general, reducing the rotation angles does not effect the cost gradients experienced. This intuition is confirmed by the numerical results displayed in the top panel of Fig. 4. Here, we randomly initialize the parameters by randomly choosing $\tilde{\theta}_l^i$ in the range $[0, 2\pi]$. We find that the cost partial derivatives for different r values perfectly overlap in this case, i.e., for a random initialization, restricting the ansatz to a

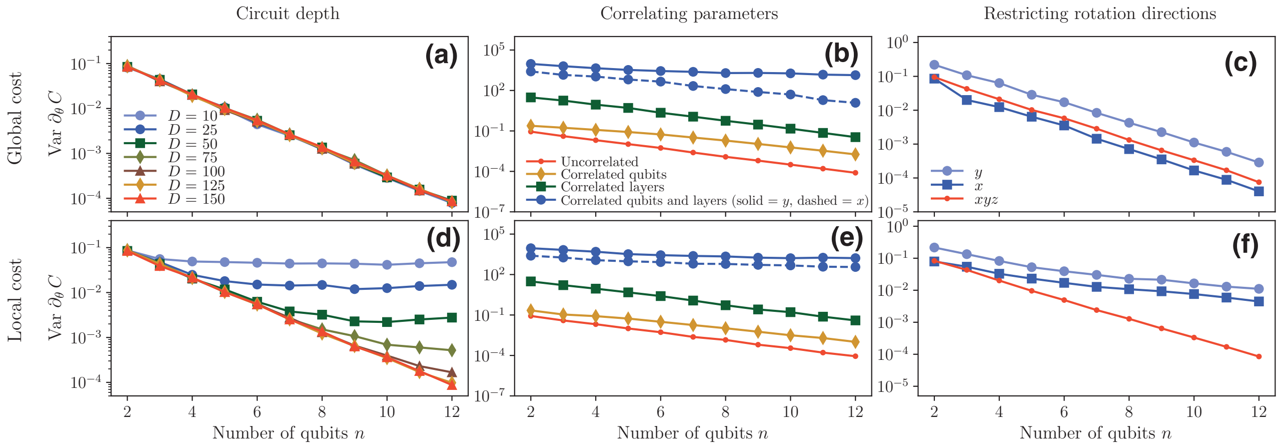

FIG. 3. Partial-derivative scalings for different expressibilities. The variance in the partial derivative of a global cost with $H_G = \bigotimes_{i=1}^{n} \sigma_i^z$ (top) and 2-local cost with $H_L = \sigma_1^z \sigma_2^z$ (bottom) as a function of the number of qubits n. In both cases $\rho = |\psi_0\rangle\langle\psi_0|^{\otimes n}$ where $|\psi_0\rangle = \exp[-i(\pi/8)\sigma_Y]|0\rangle$. In the left panel, we vary the circuit depth D of a hardware-efficient ansatz. In the middle (right) panel, we consider the effect of correlating parameters (restricting the directions of rotation) of a hardware-efficient ansatz with $D = 150$, with the choices of correlations (rotations) indicated in the figure legend. In all cases, the derivative is taken with respect to θ_1^1, the rotation angle of the first qubit in the first layer, and the variance is taken over an ensemble of 1000 unitaries.

limited range of rotation angles does not change the partial derivatives observed.

On the other hand, if the parameters are initialized close to the solution, varying r has a substantial effect on the observed partial derivatives for local costs and a reduced effect for global costs. This is seen in Figs. 4(b) and 4(c), where we initialize to identity, i.e., pick $\tilde{\theta}_l^i = 0$ for all i, which is close to the solution for this simple problem. In this case, for r close to 1 (as shown in red and yellow), the variance in the partial derivative again vanishes exponentially with n. However, for small angle ranges, $r \lessapprox 0.1$, as shown in blue, we find that the partial derivative of a local cost ceases to exhibit an exponential scaling. To some degree, a similar effect is displayed for global costs; however, the effect is reduced and is only visible in the data here for $r \approx 0.025$.

This change in partial-derivative scaling for small r for initializations close to the solution is plausibly explained by the fact that the global minimum of costs exhibiting barren plateaus tend to sit within a steep and narrow gorge [28], as sketched in Fig. 1(c). By initializing close to the solution, we are likely to be initializing within the narrow gorge. In this case, when r is close to 1, the ansatz still explores the entire cost landscape and therefore the variance in the partial derivative will be unchanged. However, for smaller r, the ansatz is constrained to the region around the narrow gorge itself and hence a larger variance in partial derivatives is observed.

5. Outlook for ansatz design

Figure 3 suggests that reducing the depth of a circuit and correlating parameters are the most effective strategies for amplifying the observed cost gradients. However, the optimal solution, of course, may not lie within a shallow or highly correlated ansatz. When deep and/or uncorrelated circuits are required, as is expected to be the case for many problems of interest, then a perturbative strategy may instead be effective. That is, one could start the variational algorithm using a shallow highly correlated ansatz and as the cost is iteratively minimized, gradually grow the ansatz [8,15,38] and decorrelate the parameters [37].

Restriction of the angle range also appears to provide an effective strategy for increasing cost gradients, but for it to be practical, it is necessary to initialize close to the solution. This, of course, requires either prior knowledge of an approximate solution to the problem at hand or an effective pretraining strategy to obtain such an approximate solution. The viability of either of these options warrants further investigation.

6. Correlation and tightness of bounds

In Fig. 6, we study the correlation between the cost gradients and our upper bounds. To quantify this correlation, we include the Spearman correlation coefficient [64], as well as its corresponding p value, which approximately gives the probability of uncorrelated data generating a Spearman coefficient at least as large as the one found. For local costs, we obtain a Spearman value of at least 0.9 with a p value of less than 0.05 in all cases, indicating a strong correlation between our upper bound and the actual variance in the gradient. The correlation is weaker in the case of global costs, highlighting that in the case of a global cost, expressibility is not the only phenomenon that may induce a barren plateau. This is to be expected given the results of Ref. [28], which show even very shallow and/or nonentangling circuits (i.e., highly inexpressive

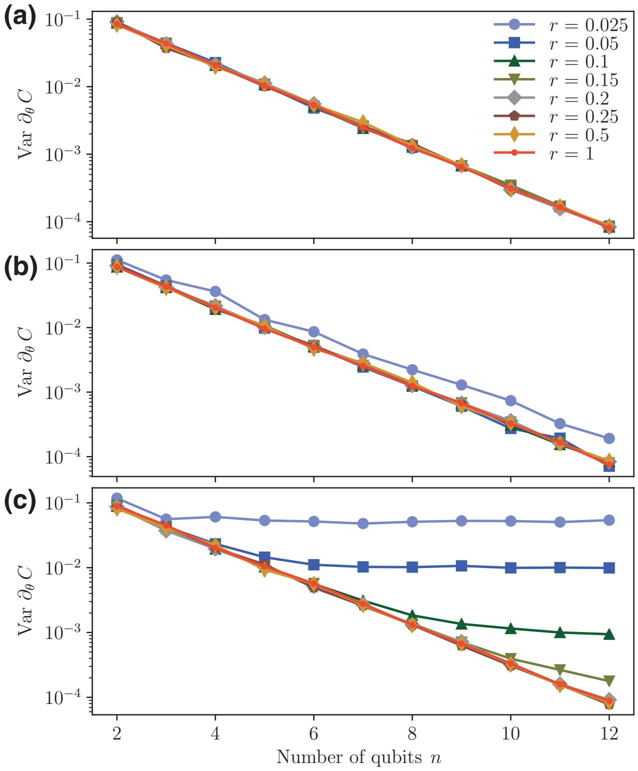

FIG. 4. Partial-derivative scalings for restricted angle ranges. The scaling of the variance in the partial derivative when the rotation angles θ_l^i are randomly chosen from the range $[\tilde{\theta}_l^i, \tilde{\theta}_l^i + 2\pi r]$, such that for $r = 1$ (red) the ansatz explores the entire solution space but for $r \ll 1$ (blue) the ansatz is constrained to exploring close to the initialization point defined by $\{\tilde{\theta}_l^i\}$. In (a), the angles $\{\tilde{\theta}_l^i\}$ are a fixed (randomly chosen) initialization point away from the solution (here, we consider a local cost but the data for a global cost are essentially unchanged). In (b) and (c), which correspond to global and local costs, respectively, the angles $\tilde{\theta}_l^i = 0$ for all l and i, which is close to the global minimum of the cost. In all cases, the derivative is taken with respect to θ_1^1, the rotation angle of the first qubit in the first layer, and the variance is taken over an ensemble of 1000 unitaries.

circuits) may exhibit barren plateaus when using global costs. For completeness, the results presented here are extended in Appendix G, where we study directly the correlation between the cost gradient and the expressibility measures ε_R^ρ and ε_L^H, similar correlations being observed.

Figure 6 additionally highlights that, as expected, the bounds are tightest for higher-expressibility ansatze but may be relatively loose for lower expressibilities. More specifically, in all cases considered here, the bounds are tight to within a couple of orders of magnitude, with the bounds tightest for ansatze that are high-depth, uncorrelated, and use the full range of rotation directions [65]. This phenomenon is more clearly demonstrated in Fig. 5, where we plot both the variance in the partial derivative of the cost and the Hamiltonian- and state-dependent expressibility bound, given in Eq. (16), as a function of the ansatz depth for the eight-qubit local cost. The bound captures the qualitative behavior of the cost gradients, which decrease with an increased circuit depth. While moderately loose at low depths, the bound becomes tight for deep circuits.

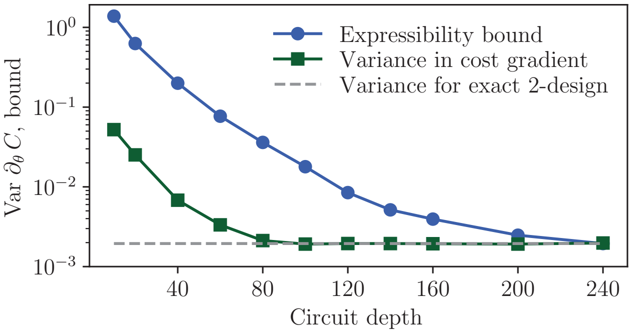

FIG. 5. Comparison of scaling of bound and gradients. The scaling of the bound on the variance in the gradient (blue), given in Eq. (16), and the variance in the partial derivative (green) as a function of the ansatz depth for $n = 8$ qubits. The dashed line indicates the predicted variance in the partial derivative for a perfect 2-design from Ref. [63]. The derivative is taken with respect to $\theta_{D/2}^1$, the rotation angle of the first qubit in the middle layer ($D/2$) and the variance is taken over an ensemble of 1000 unitaries. We choose to show the state and Hamiltonian dependent bound here, as given in Eq. (16), because as we are looking at the gradient with respect to $\theta_{D/2}^1$, this bound is tightest.

IV. DISCUSSION

In this work, we extend the well-known barren plateau result. This result was restricted to anstze that form 2-designs [27], while we extend it to arbitrary ansatze in our Theorems 1 and 2. In practice, this extension may prove to be quite useful, since many ansatze of interest are not exact 2-designs but, rather, are some approximate notion of this [56,66–68]. Our results can potentially provide useful bounds on the variance of the gradient in this realistic scenario of approximate 2-designs.

The key to our extension is to consider the expressibility of the ansatz. This can be precisely defined in terms of the distance of the ensemble of unitaries accessible by the ansatz from being a 2-design. Hence, our extension links two key properties of ansatze: their expressibility and their gradient magnitudes. Our bounds demonstrate that increasing the expressibility of an ansatz can result in smaller cost gradients. We believe that this connection is very interesting and there is certainly much more to be explored along these lines. For example, it would be interesting to connect our findings to recent results on the role of the growth of entanglement in generating barren plateaus. In particular, since highly expressive ansatze are necessarily highly entangling, our results would seem to imply those in Refs. [33,44].

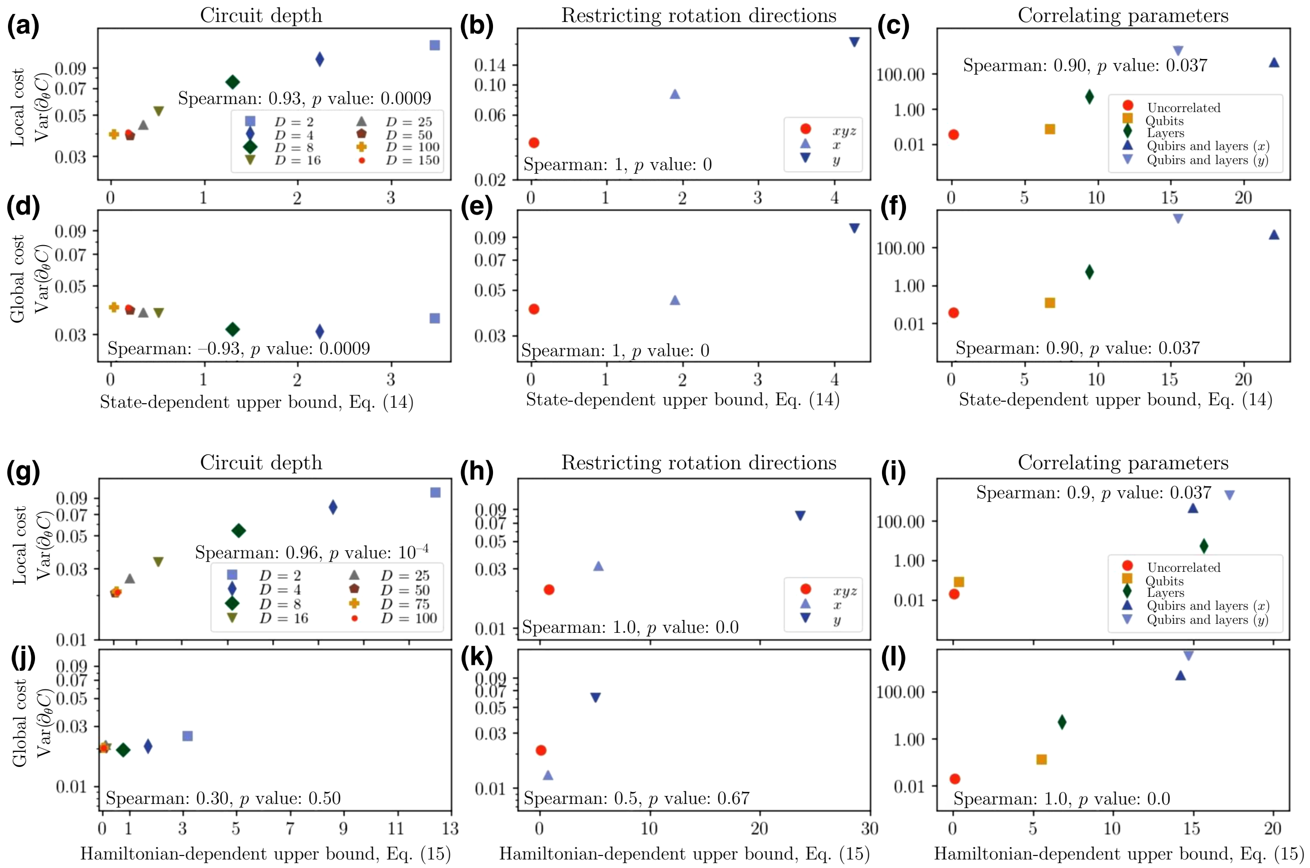

FIG. 6. Correlations between cost partial derivatives and bounds. The variance in the partial derivative of a 2-local cost with $H = \sigma_1^z\sigma_2^z$ (top) and a global cost with $H = \prod_{i=1}^n \sigma_i^z$ (bottom) as a function of (a)–(f) the state-dependent expressibility upper bound on the variance in the partial derivative specified by Eq. (14) and (g)–(l) the Hamiltonian-dependent expressibility upper bound on the variance in the partial derivative specified by Eq. (15). In both cases $\rho = |\psi_0\rangle\langle\psi_0|^{\otimes n}$ where $|\psi_0\rangle = \exp[-i(\pi/8)\sigma_Y]|0\rangle$. In the left panel, we vary the circuit depth D of a hardware-efficient ansatz. In the right (middle) panel, we consider the effect of correlating parameters (restricting the directions of rotation) of a hardware-efficient ansatz with $D = 100$, with the choices of correlations (rotations) indicated in the figure legend. In (a)–(f), the derivative is taken with respect to θ_1^1 and in (g)–(l), the derivative is taken with respect to θ_D^1. In all cases, $n = 4$ and the expressibility measures are estimated using an ensemble of 5000 unitaries.

To go beyond our bounds and look at the precise relation between expressibility and gradients, we perform extensive numerics. We consider several different strategies by which one can vary the expressibility. As highlighted in Figs. 6 and 5, we typically observe a strong correlation (especially for local cost functions) between the expressibility and the variance of the gradient. However, the bounds are not perfectly tight. This may arise from the repeated use of the triangle and Cauchy-Schwarz inequalities in the derivation (Appendix D). Thus a natural question to ask is whether our bounds can be further tightened. Another direction would be to explore the nature of barren plateaus for global costs, where the numerics suggest that the correlation between expressibility and cost gradients is weaker.

We remark that the numerical results presented here are necessarily problem specific, since they depend both on the choice in cost function and ansatz. Further work is required to ascertain the extent to which the trends observed here are universally observed. In particular, it would be valuable to investigate whether any analytic results can be obtained to support them.

Nevertheless, there are several interesting trends shown in our numerics that even suggest potential strategies of avoiding or mitigating barren plateaus. As discussed above, correlating parameters and restricting rotation angles (especially when initializing near the solution) are two strategies that significantly mitigated barren plateaus in our numerics. Further exploring these and other strategies will be an important direction for future research.

ACKNOWLEDGMENTS

Z.H. and P.J.C. were supported by the Los Alamos National Laboratory (LANL) ASC Beyond Moore’s Law project. K.S. was supported by the Laboratory Directed

Research and Development (LDRD) program of LANL under Project No. 20190065DR. M.C. was initially supported by the LDRD program of LANL under Project No. 20180628ECR and also supported by the Center for Nonlinear Studies at LANL. This work was also supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research, under the Accelerated Research in Quantum Computing (ARQC) program.

APPENDIX A: PRELIMINARIES

We begin by reviewing some definitions and prior results relevant for the rest of the appendices. We then provide proofs for the main results and theorems.

1. Operator norms

Let $\mathcal{D}(\mathcal{H})$ denote the set of density operators acting on a Hilbert space $\mathcal{H}$, i.e., those that are positive semidefinite with unit trace. Let $\mathcal{L}(\mathcal{H})$ denote the space of square linear operators acting on $\mathcal{H}$. The trace norm or Schatten 1-norm $\|\Omega\|_1$ of an operator $\Omega \in \mathcal{L}(\mathcal{H})$ is defined as $\|\Omega\|_1 := \mathrm{Tr}[|\Omega|]$, where $|\Omega| := \sqrt{\Omega^\dagger \Omega}$. More generally, the Schatten p-norm of an operator Ω can be defined as $\|\Omega\|_p = (\mathrm{Tr}[|\Omega|^p])^{1/p}$, which satisfies $\|\Omega\|_p \leqslant \|\Omega\|_q$ for $p \geqslant q$. The diamond norm of a Hermiticity-preserving linear map $\mathcal{S}_A$ is defined as

$$\|\mathcal{S}_A\|_\diamond = \sup_n \sup_{\Omega_{AB}\neq 0} \frac{\|(\mathcal{S}_A \otimes \mathcal{I}_B^{(n)})(\Omega_{AB})\|_1}{\|\Omega_{AB}\|_1}, \qquad \text{(A1)}$$

where $\Omega_{AB} \in \mathcal{L}(\mathcal{H}_A \otimes \mathcal{H}_B)$ and $\mathcal{I}_B^{(n)}$ denote an identity channel acting on an n-dimensional system B. The diamond-norm distance $\|\mathcal{N} - \mathcal{M}\|_\diamond$ is a measure of the distinguishability of two quantum operations $\mathcal{N}$ and $\mathcal{M}$.

2. Properties of the Haar measure

Let $\mathcal{U}(d)$ denote the unitary group of degree $d = 2^n$. Let $d\mu_H(V) = d\mu(V)$ be the volume element of the Haar measure, where $V \in \mathcal{U}(d)$. The volume of the Haar measure is finite: $\int_{\mathcal{U}(d)} d\mu(V) < \infty$. The Haar measure is uniquely defined up to a multiplicative constant factor. Let $d\zeta(V)$ be an invariant measure. Then there exists a constant c such that $d\zeta(V) = c\,d\mu(V)$. The Haar measure is left and right invariant under the action of the unitary group of degree d, i.e., for any integrable function $g(V)$, the following holds:

$$\begin{aligned}\int_{\mathcal{U}(d)} d\mu(V) g(WV) &= \int_{\mathcal{U}(d)} d\mu(V) g(VW)\\ &= \int_{\mathcal{U}(d)} d\mu(V) g(V),\end{aligned} \qquad \text{(A2)}$$

where $W \in \mathcal{U}(d)$.

3. Symbolic integration

We recall formulas that allow for the symbolical integration with respect to the Haar measure on a unitary group [69]. For any $V \in \mathcal{U}(d)$, the following expressions are valid for the first two moments:

$$\begin{aligned}&\int d\mu(V) v_{\mathbf{ij}} v^*_{\mathbf{pk}} = \frac{\delta_{\mathbf{ip}}\delta_{\mathbf{jk}}}{d},\\ &\times \int d\mu(V) v_{\mathbf{i}_1\mathbf{j}_1} v_{\mathbf{i}_2\mathbf{j}_2} v^*_{\mathbf{i}'_1\mathbf{j}'_1} v^*_{\mathbf{i}'_2\mathbf{j}'_2}\\ &= \frac{\delta_{\mathbf{i}_1\mathbf{i}'_1}\delta_{\mathbf{i}_2\mathbf{i}'_2}\delta_{\mathbf{j}_1\mathbf{j}'_1}\delta_{\mathbf{j}_2\mathbf{j}'_2} + \delta_{\mathbf{i}_1\mathbf{i}'_2}\delta_{\mathbf{i}_2\mathbf{i}'_1}\delta_{\mathbf{j}_1\mathbf{j}'_2}\delta_{\mathbf{j}_2\mathbf{j}'_1}}{d^2-1}\\ &\quad - \frac{\delta_{\mathbf{i}_1\mathbf{i}'_1}\delta_{\mathbf{i}_2\mathbf{i}'_2}\delta_{\mathbf{j}_1\mathbf{j}'_2}\delta_{\mathbf{j}_2\mathbf{j}'_1} + \delta_{\mathbf{i}_1\mathbf{i}'_2}\delta_{\mathbf{i}_2\mathbf{i}'_1}\delta_{\mathbf{j}_1\mathbf{j}'_1}\delta_{\mathbf{j}_2\mathbf{j}'_2}}{d(d^2-1)},\end{aligned} \qquad \text{(A3)}$$

where $v_{\mathbf{ij}}$ are the matrix elements of V. Assuming $d = 2^n$, we use the notation $\mathbf{i} = (i_1, \ldots i_n)$ to denote a bit string of length n such that $i_1, i_2, \ldots, i_n \in \{0, 1\}$.

4. Useful identities

We use the following identities, which can be derived using Eq. (A3) (for a review, see Ref. [28]):

$$\int d\mu(W) \mathrm{Tr}\left[WAW^\dagger B\right] = \frac{\mathrm{Tr}[A]\mathrm{Tr}[B]}{d}, \qquad \text{(A4)}$$

$$\begin{aligned}&\int d\mu(W) \mathrm{Tr}\left[WAW^\dagger BWCW^\dagger D\right]\\ &= \frac{\mathrm{Tr}[A]\mathrm{Tr}[C]\mathrm{Tr}[BD] + \mathrm{Tr}[AC]\mathrm{Tr}[B]\mathrm{Tr}[D]}{d^2-1}\\ &\quad - \frac{\mathrm{Tr}[AC]\mathrm{Tr}[BD] + \mathrm{Tr}[A]\mathrm{Tr}[B]\mathrm{Tr}[C]\mathrm{Tr}[D]}{d\left(d^2-1\right)},\end{aligned} \qquad \text{(A5)}$$

$$\begin{aligned}&\int d\mu(W) \mathrm{Tr}\left[WAW^\dagger B\right] \mathrm{Tr}\left[WCW^\dagger D\right]\\ &= \frac{\mathrm{Tr}[A]\mathrm{Tr}[B]\mathrm{Tr}[C]\mathrm{Tr}[D] + \mathrm{Tr}[AC] + \mathrm{Tr}[BD]}{d^2-1}\\ &\quad - \frac{\mathrm{Tr}[AC]\mathrm{Tr}[B]\mathrm{Tr}[D] + \mathrm{Tr}[A]\mathrm{Tr}[C]\mathrm{Tr}[BD]}{d\left(d^2-1\right)},\end{aligned} \qquad \text{(A6)}$$

where $A, B, C,$ and D are linear operators on a d-dimensional Hilbert space.

Let $A \in \mathcal{L}(\mathcal{H})$ and $B \in \mathcal{L}(\mathcal{H}')$. Then the following identity holds:

$$\mathrm{Tr}[A]\mathrm{Tr}[B] = \mathrm{Tr}[A \otimes B]. \qquad \text{(A7)}$$

Let $A, B \in \mathcal{L}(\mathcal{H})$, where $\mathcal{H}$ is a d^2-dimensional Hilbert space. Then, from Eq. (A3), we derive the following integral:

$$\begin{aligned}\int d\mu(U)\mathrm{Tr}[AU^{\otimes 2}BU^{\dagger\otimes 2}] &= \frac{\mathrm{Tr}[A]\mathrm{Tr}[B]+\mathrm{Tr}[AW]\mathrm{Tr}[BW]}{d^2-1}\\ &\quad -\frac{\mathrm{Tr}[AW]\mathrm{Tr}[B]+\mathrm{Tr}[A]\mathrm{Tr}[BW]}{d(d^2-1)},\end{aligned} \tag{A8}$$

where W is the subsystem swap operator, i.e., $W|i\rangle|j\rangle = |j\rangle|i\rangle$.

APPENDIX B: DEFINITIONS OF EXPRESSIBILITY

In broad terms, a parametrized quantum circuit can be considered expressive if the circuit can be used to uniformly explore the unitary group $\mathcal{U}(d)$. Thus, the expressibility of a circuit can be defined in terms of the following superoperator:

$$\begin{aligned}\mathcal{A}^{(t)}_{\mathbb{U}}(\cdot) := \int_{\mathcal{U}(d)} d\mu(V)V^{\otimes t}(\,\cdot\,)(V^{\dagger})^{\otimes t}\\ -\int_{\mathbb{U}} dU\, U^{\otimes t}(\,\cdot\,)(U^{\dagger})^{\otimes t},\end{aligned} \tag{B1}$$

where $d\mu(V)$ is the volume element of the Haar measure and dU is the volume element corresponding to the uniform distribution over $\mathbb{U}$. If $\mathcal{A}^{(t)}_{\mathbb{U}}(X) = 0$ for all operators X, then the averaging over elements of $\mathbb{U}$ agrees with averaging over the Haar distribution up to the tth moment. In this case, $\mathbb{U}$ is said to form a *t-design*. For our purposes, it suffices to consider the behavior of $\mathcal{A}^{(t)}_{\mathbb{U}}$ for $t = 2$. Henceforth, we drop the t superscript and denote $\mathcal{A}^{(2)}_{\mathbb{U}}(\cdot)$ as $\mathcal{A}_{\mathbb{U}}(\cdot)$. In the context of minimizing a generic cost C of the form specified by Eq. (1), we are interested in the quantities

$$\varepsilon^{\rho}_{\mathbb{U}} := ||\mathcal{A}_{\mathbb{U}}(\rho^{\otimes 2})||_2, \tag{B2}$$

$$\varepsilon^{H}_{\mathbb{U}} := ||\mathcal{A}_{\mathbb{U}}(H^{\otimes 2})||_2. \tag{B3}$$

The quantities $\varepsilon^{\rho}_{\mathbb{U}}$ and $\varepsilon^{H}_{\mathbb{U}}$ may be more readily computed by relating them to a generalization of the frame potential. To demonstrate how, let us first recall that the frame potential [48,53] of an ensemble $\mathbb{U}$ may be defined as

$$\mathcal{F}_{\mathbb{U}} := \int_{\mathbb{U}}\int_{\mathbb{U}} dUdV|\langle 0|(UV^{\dagger})|0\rangle|^4, \tag{B4}$$

where dU and dV are volume elements corresponding to the distribution over $\mathbb{U}$. We then note that the quantity $||\mathcal{A}_{\mathbb{U}}(|0\rangle\langle 0|)||_2^2$ can be rewritten in terms of $\mathcal{F}_{\mathbb{U}}$ as follows:

$$\begin{aligned}||\mathcal{A}_{\mathbb{U}}(|0\rangle\langle 0|)||_2^2 &= \left|\left|\int_{\mathcal{U}(d)} d\mu(V)\; V^{\otimes 2}(|0\rangle\langle 0|)(V^{\dagger})^{\otimes 2} - \int_{\mathbb{U}} dU\, U^{\otimes 2}(|0\rangle\langle 0|)(U^{\dagger})^{\otimes 2}\right|\right|_2^2\\ &= \int_{\mathbb{U}}\int_{\mathbb{U}} dUdV|\langle 0|(UV^{\dagger})|0\rangle|^4 - 2\int_{\mathcal{U}(d)}\int_{\mathbb{U}} d\mu(V)dU|\langle 0|(UV^{\dagger})|0\rangle|^4\\ &\quad + \int_{\mathcal{U}(d)}\int_{\mathcal{U}(d)} d\mu(U)d\mu(V)|\langle 0|(UV^{\dagger})|0\rangle|^4\\ &= \int_{\mathbb{U}}\int_{\mathbb{U}} dUdV|\langle 0|(UV^{\dagger})|0\rangle|^4 - \int_{\mathcal{U}(d)}\int_{\mathcal{U}(d)} d\mu(U)d\mu(V)|\langle 0|(UV^{\dagger})|0\rangle|^4\\ &= \mathcal{F}_{\mathbb{U}} - \mathcal{F}_{\mathrm{Haar}},\end{aligned} \tag{B5}$$

where we use the left and right invariance of the Haar measure, i.e., Eq. (A2), as in Ref. [53], and where we define

$$\begin{aligned}\mathcal{F}_{\mathrm{Haar}} &:= \int_{\mathcal{U}(d)}\int_{\mathcal{U}(d)} d\mu(U)d\mu(V)|\langle 0|(UV^{\dagger})|0\rangle|^4\\ &= \frac{1}{(2^n+1)2^{n-1}}.\end{aligned} \tag{B6}$$

In the context of the expressibility of a VQA, we are interested in the more general quantity $||\mathcal{A}_{\mathbb{U}}(X^{\otimes 2})||_2$, where X is a quantum state ρ or Hamiltonian H. Following the same approach as in Eq. (B5), we note that $||\mathcal{A}_{\mathbb{U}}(X^{\otimes 2})||_2$ can be rewritten as

$$||\mathcal{A}_{\mathbb{U}}(X^{\otimes 2})||_2 = \sqrt{\mathcal{F}^{(X)}_{\mathbb{U}} - \mathcal{F}^{(X)}_{\mathrm{Haar}}}, \tag{B7}$$

where we define the operator-dependent frame-potential as

$$\mathcal{F}^{(X)}_{\mathbb{U}} := \int_{\mathbb{U}}\int_{\mathbb{U}} dUdV\mathrm{Tr}[XU^{\dagger}VXV^{\dagger}U]^2 \tag{B8}$$

and

$$\mathcal{F}^{(X)}_{\text{Haar}} := \int_{\mathcal{U}(d)}\int_{\mathcal{U}(d)} d\mu(U)d\mu(V)\text{Tr}[XU^{\dagger}VXV^{\dagger}U]^2. \quad \text{(B9)}$$

The latter can be evaluated using Eq. (A6) to give

$$\mathcal{F}^{(X)}_{\text{Haar}} = \frac{\text{Tr}[X]^4 + \text{Tr}[X^2]^2}{2^{2n}-1} - \frac{2\text{Tr}[X^2]\text{Tr}[X]^2}{2^n(2^{2n}-1)}. \quad \text{(B10)}$$

Thus our expressibility measures can be related to state- and Hamiltonian-dependent frame potentials via

$$\varepsilon^{\rho}_{\mathbb{U}} := ||\mathcal{A}_{\mathbb{U}}(\rho^{\otimes 2})||_2 = \sqrt{\mathcal{F}^{(\rho)}_{\mathbb{U}} - \mathcal{F}^{(\rho)}_{\text{Haar}}}, \quad \text{(B11)}$$

$$\varepsilon^{H}_{\mathbb{U}} := ||\mathcal{A}_{\mathbb{U}}(H^{\otimes 2})||_2 = \sqrt{\mathcal{F}^{(H)}_{\mathbb{U}} - \mathcal{F}^{(H)}_{\text{Haar}}}. \quad \text{(B12)}$$

We use these expressions to evaluate the expressibility of different ansatze in Appendix G.

APPENDIX C: PROOF FOR EQ. (8)

For a random layered parametrized ansatz of the form Eqs. (2) and Eqs. (11)–(12), and the generic cost defined in Eq. (1), we now show that $\langle \partial_k C\rangle_{\mathbb{U}} = 0$ for all k and therefore that the cost landscape is unbiased.

To do so, let us first note that the cost function can be expressed as

$$C = \text{Tr}[U_k(\theta_k)\widetilde{\rho}U_k(\theta_k)^{\dagger}\widetilde{H}], \quad \text{(C1)}$$

where we introduce the shorthand

$$\widetilde{\rho} = W_k\left(\prod_{j=1}^{k-1}U_j(\theta_j)W_j\right)\rho\left(\prod_{j=1}^{k-1}U_j(\theta_j)W_j\right)^{\dagger}W_k^{\dagger}, \quad \text{(C2)}$$

$$\widetilde{H} = U_L(\boldsymbol{\theta})^{\dagger}HU_L(\boldsymbol{\theta}). \quad \text{(C3)}$$

This rewriting emphasizes the dependence of C on $U_k(\theta_k)$, the rotation with respect to which we are taking the partial derivative, by associating U_L with the Hamiltonian H and $\left(\prod_{j=1}^{k-1}U_j(\theta_j)W_j\right)$ with ρ.

It follows that

$$\begin{aligned}\partial_k C = &-i\text{Tr}[V_kU_k(\theta_k)\widetilde{\rho}U_k(\theta_k)^{\dagger}\widetilde{H}]\\ &+i\text{Tr}[U_k(\theta_k)\widetilde{\rho}U_k(\theta_k)^{\dagger}V_k\widetilde{H}]\end{aligned} \quad \text{(C4)}$$

$$\begin{aligned}= &-i\text{Tr}[V_k(\cos(\theta_k) - i\sin(\theta_k)V_k)\widetilde{\rho}(\cos(\theta_k)\\ &+ i\sin(\theta_k)V_k)\widetilde{H}] + i\text{Tr}[(\cos(\theta_k)\\ &- i\sin(\theta_k)V_k)\widetilde{\rho}(\cos(\theta_k) + i\sin(\theta_k)V_k)V_k\widetilde{H}]\end{aligned} \quad \text{(C5)}$$

$$\begin{aligned}= &-i\left((\cos(\theta_k)^2 - \sin(\theta_k)^2)(\text{Tr}[V_k\widetilde{\rho}\widetilde{H}] - \text{Tr}[\widetilde{\rho}V_k\widetilde{H}])\right.\\ &\left.+\, i\sin(2\theta_k)(\text{Tr}[V_k\widetilde{\rho}V_k\widetilde{H}] - \text{Tr}[\widetilde{\rho}\widetilde{H}])\right).\end{aligned} \quad \text{(C6)}$$

Since $\int_0^{2\pi}\sin(2\theta_k) = 0$ and $\int_0^{2\pi}\cos(\theta_k)^2 = \int_0^{2\pi}\sin(\theta_k)^2$, uniform averaging of $\partial_k C$ over θ_k leads to

$$\frac{1}{2\pi}\int_0^{2\pi} d\theta_k\partial_k C = 0, \quad \text{(C7)}$$

which implies that $\langle\partial_k C\rangle_{\mathbb{U}} = 0$.

APPENDIX D: VARIANCE OF THE PARTIAL-DERIVATIVE DERIVATION

For a random layered parametrized ansatz of the form Eqs. (2) and Eqs. (11)–(12), and the generic cost defined in Eq. (1), then since

$$\partial_k U(\boldsymbol{\theta}) = -iU_LV_kU_R \quad \text{(D1)}$$

where U_L and U_R are defined in Eq. (12), it follows that the partial derivative of the cost can be written as

$$\partial_k C := \frac{\partial C}{\partial\theta_k} = i\text{Tr}[U_R\rho U_R^{\dagger}[V_k, U_L^{\dagger}HU_L]]. \quad \text{(D2)}$$

Since the average derivative of the cost vanishes, as discussed in Appendix C, its variance is given by

$$\text{Var}\ \partial_k C = \langle(\partial_k C)^2\rangle_{\mathbb{U}}. \quad \text{(D3)}$$

Equations (D2) and (D3) provide the starting point to derive the bounds Eqs. (14)–(16) and Eqs. (19)–(21).

1. Bound in Eq. (14)

Note that two different ensembles $\mathbb{U}_L$ and $\mathbb{U}_R$ can be generated using $U_L(\boldsymbol{\theta})$ and $U_R(\boldsymbol{\theta})$, respectively, as defined in Eq. (11). Let dU_L and dU_R denote volume elements corresponding to distributions over $\mathbb{U}_L$ and $\mathbb{U}_R$, respectively. Since $\mathbb{U}_L$ and $\mathbb{U}_R$ are independent, from the definition of dU and from Eq. (11), we obtain that $dU = dU_LdU_R$.

Then, by substituting Eq. (D2) into Eq. (D3) and using Eq. (A7), we obtain

$$\text{Var}\partial_k C = -\int_{\mathbb{U}_L}dU_L\int_{\mathbb{U}_R}dU_R\text{Tr}[U_R^{\otimes 2}\rho^{\otimes 2}U_R^{\dagger\otimes 2}X_{Lk}^{\otimes 2}], \quad \text{(D4)}$$

where

$$X_{Lk} := [V_k, U_L^{\dagger}HU_L]. \quad \text{(D5)}$$

Next, we substitute in $\mathcal{A}_R(\rho^{\otimes 2})$ to give

$$\begin{aligned}\mathrm{Var}\partial_k C &= -\int_{\mathbb{U}_L} dU_L \int_{\mathcal{U}(d)} d\mu(U)\mathrm{Tr}[U_{\mathrm{Haar}}^{\otimes 2}\rho^{\otimes 2}U_{\mathrm{Haar}}^{\dagger\otimes 2}X_{Lk}^{\otimes 2}] \\ &\quad + \int_{\mathbb{U}_L} dU_L \mathrm{Tr}[\mathcal{A}_R(\rho^{\otimes 2})X_{Lk}^{\otimes 2}] \\ &= \mathrm{Var}_R\partial_k C + \int_{\mathbb{U}_L} dU_L \mathrm{Tr}[\mathcal{A}_R(\rho^{\otimes 2})X_{Lk}^{\otimes 2}], \qquad \text{(D6)}\end{aligned}$$

where in the second line we use the explicit definition of $\mathrm{Var}_R\partial_k C$, the variance in the partial derivative of the cost when $\mathbb{U}_\mathbb{R}$ forms a 2-design, i.e.,

$$\begin{aligned}\mathrm{Var}_R\partial_k C := -\int_{\mathbb{U}_L} dU_L \int_{\mathcal{U}(d)} d\mu(U)\mathrm{Tr} \\ \times [U_{\mathrm{Haar}}^{\otimes 2}\rho^{\otimes 2}U_{\mathrm{Haar}}^{\dagger\otimes 2}X_{Lk}^{\otimes 2}]. \qquad \text{(D7)}\end{aligned}$$

Rearranging, we are left with

$$|\mathrm{Var}\ \partial_k C - \mathrm{Var}_R\partial_k C| \leqslant \left|\int_{\mathbb{U}_L} dU_L \mathrm{Tr}[\mathcal{A}_R(\rho^{\otimes 2})X_{Lk}^{\otimes 2}]\right|, \qquad \text{(D8)}$$

which on using the triangle inequality followed by the Cauchy-Schwarz inequality reduces to

$$\begin{aligned}|\mathrm{Var}\ \partial_k C - \mathrm{Var}_R\partial_k C| &\leqslant \int_{\mathbb{U}_L} dU_L |\mathrm{Tr}[\mathcal{A}_R(\rho^{\otimes 2})X_{Lk}^{\otimes 2}]| \\ &\leqslant \int_{\mathbb{U}_L} dU_L ||X_{Lk}^{\otimes 2}||_2 ||\mathcal{A}_R(\rho^{\otimes 2})||_2. \qquad \text{(D9)}\end{aligned}$$

The term $||X_{Lk}^{\otimes 2}||_2$ can be bounded as follows. First, we note that $X_{Lk}^\dagger = -X_{Lk}$, which implies that

$$\begin{aligned}||X_{Lk}^{\otimes 2}||_2 &= \sqrt{\mathrm{Tr}[X_{Lk}^{\otimes 2}X_{Lk}^{\otimes 2}]} = \sqrt{\mathrm{Tr}[X_{Lk}^2 \otimes X_{Lk}^2]} \\ &= |\mathrm{Tr}[X_{Lk}^2]| = |\mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2]|. \qquad \text{(D10)}\end{aligned}$$

Let $A = V_k$ and $B = U_L^\dagger H U_L$. Since A and B are Hermitian, from the triangle inequality and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}|\mathrm{Tr}[[A,B]^2]| &= 2|\mathrm{Tr}[ABAB] - \mathrm{Tr}[A^2B^2]| \leqslant 2[|\mathrm{Tr}[ABAB]| \\ &\quad + |\mathrm{Tr}[B^2]|] \leqslant 2\sqrt{\mathrm{Tr}[ABAABA]\mathrm{Tr}[B^2]} \\ &\quad + 2|\mathrm{Tr}[B^2]| = 4|\mathrm{Tr}[B^2]|. \qquad \text{(D11)}\end{aligned}$$

Therefore, we find that

$$||X_{Lk}^{\otimes 2}||_2 \leqslant 4\mathrm{Tr}[(U_L^\dagger H U_L)^2] = 4\mathrm{Tr}[H^2] = 4||H||_2^2. \qquad \text{(D12)}$$

Hence the bound takes the form

$$\begin{aligned}|\mathrm{Var}\ \partial_k C - \mathrm{Var}_R\partial_k C| &\leqslant 4\int_{\mathbb{U}_L} dU_L ||\mathcal{A}_R(\rho^{\otimes 2})||_2 ||H||_2^2 \\ &= 4||\mathcal{A}_R(\rho^{\otimes 2})||_2 ||H||_2^2, \qquad \text{(D13)}\end{aligned}$$

which completes the proof.

a. Extension to generalized cost

This result can be further extended to cost functions of the following form:

$$C(\boldsymbol{\theta}) = \sum_m \mathrm{Tr}[H_m U(\boldsymbol{\theta})\rho_m U(\boldsymbol{\theta})^\dagger], \qquad \text{(D14)}$$

for which the derivative with respect to the parameter θ_k can be written as

$$\partial_k C = i\sum_m \mathrm{Tr}[U_R\rho_m U_R^\dagger[V_k, U_L^\dagger H_m U_L]]. \qquad \text{(D15)}$$

Therefore, from Eq. (D7), it follows that

$$\begin{aligned}\mathrm{Var}\partial_k C = -\sum_{m,n}\int_{\mathbb{U}_L} dU_L \int_{\mathbb{U}_R} dU_R \mathrm{Tr} \\ \times [U_R^{\otimes 2}(\rho_m \otimes \rho_n)(U_R^\dagger)^{\otimes 2}X_{Lk}^m \otimes X_{Lk}^n], \qquad \text{(D16)}\end{aligned}$$

where X_{Lk}^m is defined in Eq. (D5) with $H = H_m$.

After substituting $\mathcal{A}_R(\rho_j \otimes \rho_k)$, we obtain

$$\begin{aligned}\mathrm{Var}\partial_k C = \mathrm{Var}_R\partial_k C + \sum_{m,n}\int_{\mathbb{U}_L} dU_L \mathrm{Tr} \\ \times [\mathcal{A}_R(\rho_m \otimes \rho_n)(X_{Lk}^m \otimes X_{Lk}^n)], \qquad \text{(D17)}\end{aligned}$$

which implies that

$$\begin{aligned}&|\mathrm{Var}\ \partial_k C - \mathrm{Var}_R\partial_k C| \\ &\leqslant \sum_{m,n}\int_{\mathbb{U}_L} dU_L |\mathrm{Tr}[\mathcal{A}_R(\rho_m \otimes \rho_n)(X_{Lk}^m \otimes X_{Lk}^n)]| \qquad \text{(D18)} \\ &\leqslant \sum_{m,n}\int_{\mathbb{U}_L} dU_L \|\mathcal{A}_R(\rho_m \otimes \rho_n)\|_2 \|(X_{Lk}^m \otimes X_{Lk}^n)\|_2 \qquad \text{(D19)} \\ &\leqslant \sum_{m,n} \|\mathcal{A}_R(\rho_m \otimes \rho_n)\|_2 \sqrt{\mathrm{Tr}[(X_{Lk}^m)^2]\mathrm{Tr}[(X_{Lk}^n)^2]} \qquad \text{(D20)} \\ &\leqslant 4\sum_{m,n} \|\mathcal{A}_R(\rho_m \otimes \rho_n)\|_2 \|H_m\|_2 \|H_n\|_2, \qquad \text{(D21)}\end{aligned}$$

where we use steps similar to those used in deriving Eqs. (D10)–(D13).

2. Bound in Eq. (15)

Substituting Eq. (D2) into Eq. (D3) and using Eq. (A7) and the cyclicity of the trace operation, we find that

$$\mathrm{Var}\partial_k C = \int_{\mathbb{U}_L} dU_L \int_{\mathbb{U}_R} dU_R \mathrm{Tr}[U_L^{\dagger\otimes 2} H^{\otimes 2} U_L^{\otimes 2} Y_{Rk}^{\otimes 2}], \tag{D22}$$

where $Y_{Rk} := [U_R \rho U_R^\dagger, V_k]$. The rest of the derivation proceeds in the same manner as for the bound in Eq. (14).

a. Extension to generalized cost

Similar to Eq. (D21), the bound in Eq. (15) can be extended for the cost functions of the form in Eq. (D14). In particular, we find that

$$\begin{aligned} &|\mathrm{Var}\partial_k C - \mathrm{Var}_L \partial_k C| \\ &\quad \leqslant 4 \sum_{m,n} \|\mathcal{A}_L(H_m \otimes H_n)\|_2 \|\rho_m\|_2 \|\rho_n\|_2 . \end{aligned} \tag{D23}$$

3. Bound in Eq. (16)

To derive Eq. (16), we start by substituting Eq. (D2) into Eq. (D3) and using Eq. (A7) and the cyclicity of the trace operation to find that

$$\begin{aligned} \mathrm{Var}\partial_k C = &-\int_{\mathbb{U}_L} dU_L \int_{\mathbb{U}_R} dU_R \mathrm{Tr}[\rho_R^{\otimes 2}(V_k^{\otimes 2} H_L^{\otimes 2} + H_L^{\otimes 2} V_k^{\otimes 2} \\ &- 2(V_k \otimes \mathbb{1}) H_L^{\otimes 2}(\mathbb{1} \otimes V_k))], \end{aligned} \tag{D24}$$

where we introduce the shorthand $\rho_R := U_R \rho U_R^\dagger$ and $H_L := U_L^\dagger H U_L$. Next, we substitute in $\mathcal{A}_L(H^{\otimes 2})$ and $\mathcal{A}_R(\rho^{\otimes 2})$ to find that the variance is given by

$$\mathrm{Var}\partial_k C = \mathrm{Var}_{L,R} \partial_k C - \mathrm{Tr}[\mathcal{A}_R(\rho^{\otimes 2}) Z_{Lk}] + I_1 + I_2. \tag{D25}$$

Here, we define

$$\begin{aligned} Z_{xk} := [&V_k^{\otimes 2} \mathcal{A}_x(\omega_x) + \mathcal{A}_x(\omega_x) V_k^{\otimes 2} \\ &- 2(V_k \otimes \mathbb{1}) \mathcal{A}_x(\omega_x)(\mathbb{1} \otimes V_k)], \end{aligned} \tag{D26}$$

for $x = L$ and $x = R$, and where $\omega_R = \rho$ and $\omega_L = H$. The integrals I_1 and I_2 are given by

$$\begin{aligned} I_1 &= \int_{\mathcal{U}(d)} d\mu(U) \mathrm{Tr}[Z_{Lk} \tilde{\rho}^{\otimes 2}] \\ I_2 &= \int_{\mathcal{U}(d)} d\mu(U) \mathrm{Tr}[\mathcal{A}_R(\rho^{\otimes 2})(V_k^{\otimes 2} \tilde{H}^{\otimes 2} + \tilde{H}^{\otimes 2} V_k^{\otimes 2} \\ &\quad - 2(V_k \otimes \mathbb{1}) \tilde{H}^{\otimes 2}(\mathbb{1} \otimes V_k)], \end{aligned} \tag{D27}$$

with $\tilde{\rho} = U\rho U^\dagger$ and $\tilde{H} = U^\dagger H U$.

The integrals I_1 and I_2 can be evaluated using Eq. (A8) as follows:

$$\begin{aligned} I_1 &= \frac{1}{d^2-1} \mathrm{Tr}[Z_{Lk} W] \mathrm{Tr}[\rho^2] - \frac{1}{d(d^2-1)} \mathrm{Tr}[Z_{Lk} W], \\ I_2 &= \frac{1}{d^2-1} \mathrm{Tr}[Z_{Rk} W] \mathrm{Tr}[H^2] - \frac{1}{d(d^2-1)} \mathrm{Tr}[Z_{Rk} W] \mathrm{Tr}[H]^2, \end{aligned} \tag{D28}$$

where we use the fact that $\mathrm{Tr}[Z_{Lk}] = \mathrm{Tr}[Z_{Rk}] = 0$, $\mathrm{Tr}[\rho^{\otimes 2} W] = \mathrm{Tr}[\rho^2]$, and $\mathrm{Tr}[H^{\otimes 2} W] = \mathrm{Tr}[H^2]$.

Substituting these integrals, given in Eq. (D28), back into Eq. (D25) and then using the triangle inequality yields

$$\begin{aligned} &|\mathrm{Var}\ \partial_k C - \mathrm{Var}_{R,L} \partial_k C| \\ &\quad \leqslant \frac{1}{d^2-1}((\mathrm{Tr}[\rho^2] - 1/d)|\mathrm{Tr}[Z_{Lk} W]| \\ &\qquad + (\mathrm{Tr}[H^2] - \mathrm{Tr}[H]^2/d)|\mathrm{Tr}[Z_{Rk} W]|) \\ &\qquad + |\mathrm{Tr}[\mathcal{A}_R(\rho^{\otimes 2}) Z_{L,k}]|. \end{aligned} \tag{D29}$$

Using Cauchy-Schwarz, this reduces to

$$\begin{aligned} |\mathrm{Var}\ \partial_k C - \mathrm{Var}_{R,L} \partial_k C| &\leqslant \frac{d}{d^2-1}[(||\rho||_2^2 - 1/d)||Z_{Lk}||_2 \\ &+ (||H||_2^2 - \mathrm{Tr}[H]^2/d)||Z_{Rk}||_2] + ||\mathcal{A}_R(\rho^{\otimes 2})||_2 ||Z_{Lk}||_2, \end{aligned} \tag{D30}$$

where we use $||W||_2 = d$. Finally, by expanding $||Z_{xk}||_2$, using the triangle inequality and the fact that $V_k^2 = \mathbb{1}$, we find that

$$||Z_{xk}||_2 \leqslant 4||\mathcal{A}_x(\omega_x)||_2, \tag{D31}$$

for $x = L$ and $x = R$, and where $\omega_R = \rho$ and $\omega_L = H$. Thus we are left with

$$\begin{aligned} |\mathrm{Var}\partial_k C - \mathrm{Var}_{R,L} \partial_k C| &\leqslant 4||\mathcal{A}_R(\rho^{\otimes 2})||_2 ||\mathcal{A}_L(H^{\otimes 2})||_2 + \frac{2^{n+2}}{2^{2n}-1} \\ &\times \left[||\mathcal{A}_R(\rho^{\otimes 2})||_2 \left(||H||_2^2 - \frac{1}{d}\mathrm{Tr}[H]^2\right) + ||\mathcal{A}_L(H^{\otimes 2})||_2 \left(||\rho||_2^2 - \frac{1}{d}\right)\right]. \end{aligned} \tag{D32}$$

a. Extension to generalized cost

Similar to Eqs. (D21) and (D23), the bound in Eq. (16) can be extended for the cost functions of the form in Eq. (D14). In particular, we find that

$$|\mathrm{Var}\partial_k C - \mathrm{Var}_{R,L}\partial_k C| \leqslant 4\sum_{m,n} ||\mathcal{A}_R(\rho_m \otimes \rho_n)||_2 ||\mathcal{A}_L(H_m \otimes H_n)||_2 + \frac{2^{n+2}}{2^{2n}-1}\sum_{m,n}$$
$$\times\left[||\mathcal{A}_R(\rho_m \otimes \rho_n)||_2\left(\mathrm{Tr}[H_m H_n] - \frac{1}{d}\mathrm{Tr}[H_m]\mathrm{Tr}[H_n]\right) + ||\mathcal{A}_L(H_m \otimes H_n)||_2\left(\mathrm{Tr}[\rho_m\rho_n] - \frac{1}{d}\right)\right]. \tag{D33}$$

4. Reformulating bounds using the diamond norm

Here, we derive bounds Eqs. (19)–(21), in which the expressibility is quantified in terms of the diamond norm. This is a natural alternative way of formulating the bounds, since the diamond norm is an operationally meaningful measure of the distinguishability of two quantum operations that is often used to define ε-approximate t-designs.

To derive Eq. (19), we start with Eq. (D9) and invoke the Hölder's inequality as follows:

$$|\mathrm{Var}\ \partial_k C - \mathrm{Var}_R \partial_k C| \leqslant \int_{\mathbb{U}_L} dU_L |\mathrm{Tr}[\mathcal{A}_R(\rho^{\otimes 2}) X_{Lk}^{\otimes 2}]| \leqslant \int_{\mathbb{U}_L} dU_L ||X_{Lk}^{\otimes 2}||_\infty ||\mathcal{A}_R(\rho^{\otimes 2})||_1. \tag{D34}$$

The term $||X_{Lk}^{\otimes 2}||_\infty$ can now be bounded as follows. Given that $X_{Lk}^\dagger = -X_{Lk}$, it follows from the unitary invariance and submultiplicativity of the infinity norm that

$$||X_{Lk}^{\otimes 2}||_\infty = (||X_{Lk}||_\infty)^2 \leqslant (2||V_k||_\infty ||U_L^\dagger H U_L||_\infty)^2 = (2||V_k||_\infty ||H||_\infty)^2 \leqslant 4||H||_\infty^2. \tag{D35}$$

We additionally note that $||\mathcal{E}(X)||_1 \leqslant \|X\|_1 ||\mathcal{E}||_\diamond$ for any channel $\mathcal{E}$ and operator X. Therefore,

$$||\mathcal{A}_R(\rho^{\otimes 2})||_1 \leqslant \|\rho\|_1 ||\mathcal{A}_{U_R}||_\diamond = ||\mathcal{A}_{U_R}||_\diamond := \varepsilon_R^\diamond. \tag{D36}$$

Thus we are now left with

$$|\mathrm{Var}\ \partial_k C - \mathrm{Var}_R \partial_k C| \leqslant 4||H||_\infty^2\ \varepsilon_R^\diamond. \tag{D37}$$

The derivation of Eq. (20) is entirely analogous.

To derive Eq. (21), we start with Eq. (D29) and again use Hölder's inequality in terms of the infinity and one norm to find

$$|\mathrm{Var}\ \partial_k C - \mathrm{Var}_{R,L}\partial_k C| \leqslant \frac{1}{d^2-1}[(||\rho||_2^2 - 1/d)||Z_{Lk}||_1 + (||H||_2^2 - \mathrm{Tr}[H]^2/d)||Z_{Rk}||_1] + ||\mathcal{A}_R(\rho^{\otimes 2})||_1 ||Z_{Lk}||_\infty, \tag{D38}$$

where we use $||W||_\infty = 1$. Finally, by expanding $||Z_{xk}||_\infty$, using the triangle inequality and the fact that $V_k^2 = \mathbb{1}$, we find that

$$||Z_{xk}||_\infty \leqslant ||V_k^{\otimes 2}\mathcal{A}_x(\omega_x)||_\infty + ||\mathcal{A}_x(\omega_x)V_k^{\otimes 2}||_\infty + 2||(V_k \otimes \mathbb{1})\mathcal{A}_x(\omega_x)(\mathbb{1}\otimes V_k)||_\infty \leqslant 2||V_k^{\otimes 2}||_\infty ||\mathcal{A}_x(\omega_x)||_\infty$$
$$+ 2||(V_k \otimes \mathbb{1})||_\infty ||\mathcal{A}_x(\omega_x)||_\infty ||(\mathbb{1}\otimes V_k)||_\infty \leqslant 4||\mathcal{A}_x(\omega_x)||_\infty, \tag{D39}$$

for $x = L$ and $x = R$. We additionally note that $||\mathcal{E}(X)||_1 \leqslant \|X\|_1 ||\mathcal{E}||_\diamond$ for any channel $\mathcal{E}$ and operator X. Therefore,

$$||\mathcal{A}_R(\rho^{\otimes 2})||_\infty \leqslant ||\mathcal{A}_R(\rho^{\otimes 2})||_1 \leqslant \varepsilon_R^\diamond, \tag{D40}$$

$$||\mathcal{A}_L(H^{\otimes 2})||_\infty \leqslant ||\mathcal{A}_L(H^{\otimes 2})||_1 \leqslant \|H\|_1 \varepsilon_L^\diamond. \tag{D41}$$

Thus we are left with

$$|\mathrm{Var}\ \partial_k C - \mathrm{Var}_{R,L}\partial_k C| \leqslant \frac{4}{d^2-1}[(||\rho||_2^2 - 1/d)||\mathcal{A}_L(H)||_\infty + (||H||_2^2 - \mathrm{Tr}[H]^2/d)||\mathcal{A}_R(\rho^{\otimes 2})||_\infty] + 4||\mathcal{A}_R(\rho)||_1||\mathcal{A}_L(H^{\otimes 2})||_\infty \tag{D42}$$

or, alternatively,

$$|\mathrm{Var}\ \partial_k C - \mathrm{Var}_{R,L}\partial_k C| \leqslant \frac{4}{d^2-1}[(||\rho||_2^2 - 1/d)\|H\|_1\varepsilon_L^\diamond + (||H||_2^2 - \mathrm{Tr}[H]^2/d)\varepsilon_R^\diamond] + 4\|H\|_1\ \varepsilon_R^\diamond\varepsilon_L^\diamond. \tag{D43}$$

APPENDIX E: VARIANCE IN PARTIAL DERIVATIVE FOR EXACT 2-DESIGNS

In this appendix, we provide the explicit expressions and the derivation of the variance in the partial derivative for a random layered parametrized ansatz of the form Eqs. (2) and Eqs. (11)–(12) and the generic cost defined in Eq. (1). These quantities have been investigated in Ref. [27]; however, only the highest-order terms in n were given. Here, we provide higher-order terms for completeness.

1. Explicit expressions

Let us denote the variance of the cost when just $\mathbb{U}_R$, just $\mathbb{U}_L$, and both $\mathbb{U}_R$ and $\mathbb{U}_L$ form 2-designs as $\mathrm{Var}_R\partial_k C$, $\mathrm{Var}_L\partial_k C$, and $\mathrm{Var}_{R,L}\partial_k C$, respectively. These variances are given by

$$\mathrm{Var}_x\partial_k C = \frac{g_x(\rho, H, U)}{2^{2n}-1}, \tag{E1}$$

where

$$g_R(\rho, H, U) = -\left(\mathrm{Tr}[\rho^2] - \frac{1}{2^n}\right)\int dU_L \mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2], \tag{E2}$$

$$g_L(\rho, H, U) = -\left(\mathrm{Tr}[H^2] - \frac{\mathrm{Tr}[H]^2}{2^n}\right)\int dU_R \mathrm{Tr}[[V_k, U_R^\dagger H U_R]^2], \tag{E3}$$

$$\begin{aligned} g_{R,L}(\rho, H, U) = -2\left(\mathrm{Tr}[\rho^2] - \frac{1}{2^n}\right)\Big\{ & \frac{1}{2^{2n}-1}[\mathrm{Tr}[V_k]^2\mathrm{Tr}[H^2] + \mathrm{Tr}[V_k^2]\mathrm{Tr}[H]^2] \\ & - \frac{1}{2^n(2^{2n}-1)}[\mathrm{Tr}[V_k^2]\mathrm{Tr}[H^2] + \mathrm{Tr}[V_k]^2\mathrm{Tr}[H]^2] - \frac{1}{2^n}\mathrm{Tr}[V_k^2]\mathrm{Tr}[H^2]\Big\}. \end{aligned} \tag{E4}$$

2. Derivation

From Eq. (D2), we have

$$\partial_k C := \frac{\partial C}{\partial\theta_k} = i\mathrm{Tr}[U_R\rho U_R^\dagger[V_k, U_L^\dagger H U_L]]. \tag{E5}$$

Since the cost gradient is unbiased, as in Eq. (8), the variance in the partial derivative is given by

$$\mathrm{Var}\partial_k C = -\int dU_L \int dU_R \mathrm{Tr}[U_R\rho U_R^\dagger[V_k, U_L^\dagger H U_L]]^2. \tag{E6}$$

Then $\mathrm{Var}_R\partial_k C$, $\mathrm{Var}_L\partial_k C$, and $\mathrm{Var}_{R,L}\partial_k C$ can be calculated by the integration in Eq. (E6) over U_R, U_L, and both U_R and U_L, respectively.

Integrating over only U_R gives

$$\begin{aligned}\mathrm{Var}_R\partial_k C &= -\frac{1}{d^2-1}\int dU_L[\mathrm{Tr}[\rho]^2\mathrm{Tr}[[V_k, U_L^\dagger H U_L]]^2 + \mathrm{Tr}[\rho^2]\mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2]] \\ &\quad + \frac{1}{d(d^2-1)}\int dU_L[\mathrm{Tr}[\rho^2]\mathrm{Tr}[[V_k, U_L^\dagger H U_L]]^2 + \mathrm{Tr}[\rho]^2\mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2]] \qquad \text{(E7)} \\ &= -\frac{1}{d^2-1}\int dU_L\mathrm{Tr}[\rho^2]\mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2] + \frac{1}{d(d^2-1)}\int dU_L\mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2] \qquad \text{(E8)} \\ &= -\left(\mathrm{Tr}[\rho^2] - \frac{1}{d}\right)\left(\frac{1}{d^2-1}\right)\int dU_L\mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2], \qquad \text{(E9)}\end{aligned}$$

where the first equality follows from Eq. (A6) and the second equality follows from the fact that the trace of a commutator is always zero.

Form the cyclicity of the trace operation and the arguments similar to Eqs. (E7) and (E8), we obtain

$$\mathrm{Var}_L\partial_k C = -\left(\mathrm{Tr}[H^2] - \frac{\mathrm{Tr}[H]^2}{d}\right)\left(\frac{1}{d^2-1}\right)\int dU_R\mathrm{Tr}[[V_k, U_R^\dagger H U_R]^2]. \qquad \text{(E10)}$$

In order to calculate $\mathrm{Var}_{R,L}\partial_k C$, we note that $\mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2]$ in Eq. (E9) can be written as

$$\mathrm{Tr}[[V_k, U_L^\dagger H U_L]^2] = 2[\mathrm{Tr}[U_L V_k U_L^\dagger H U_L V_k U_L^\dagger H] - \mathrm{Tr}[U_L V_k^2 U_L^\dagger H^2]]. \qquad \text{(E11)}$$

The integral of the first term over U_L in Eq. (E11) can be calculated using Eq. (A5) as follows:

$$\begin{aligned}&\int dU_L\mathrm{Tr}[U_L V_k U_L^\dagger H U_L V_k U_L^\dagger H] \\ &= \frac{1}{d^2-1}[\mathrm{Tr}[V_k]^2\mathrm{Tr}[H^2] + \mathrm{Tr}[V_k^2]\mathrm{Tr}[H]^2] - \frac{1}{d(d^2-1)}[\mathrm{Tr}[V_k^2]\mathrm{Tr}[H^2] + \mathrm{Tr}[V_k]^2\mathrm{Tr}[H]^2]. \qquad \text{(E12)}\end{aligned}$$

The integral of the second term in Eq. (E11) can be calculated using Eq. (A4) as follows:

$$\int dU_L\mathrm{Tr}[U_L V_k^2 U_L^\dagger H^2] = \frac{\mathrm{Tr}[V_k^2]\mathrm{Tr}[H^2]}{d}. \qquad \text{(E13)}$$

Finally, after combining everything, we obtain

$$\begin{aligned}\mathrm{Var}_{R,L}\partial_k C = &-\Big(\mathrm{Tr}[\rho^2] - \frac{1}{d}\Big)\Big(\frac{2}{d^2-1}\Big)\Big(\frac{1}{d^2-1}[\mathrm{Tr}[V_k]^2\mathrm{Tr}[H^2] + \mathrm{Tr}[V_k^2]\mathrm{Tr}[H]^2] \\ &- \frac{1}{d(d^2-1)}[\mathrm{Tr}[V_k^2]\mathrm{Tr}[H^2] + \mathrm{Tr}[V_k]^2\mathrm{Tr}[H]^2] - \frac{1}{d}\mathrm{Tr}[V_k^2]\mathrm{Tr}[H^2]\Big). \qquad \text{(E14)}\end{aligned}$$

APPENDIX F: CONCENTRATION OF MEASURE

In Ref. [33] it has been shown that for ansatze where the reduced state on the measured qubits obeys a volume law, typical local cost function values concentrate exponentially fast in n to its mean. This result has been complemented by a proof that for ansatze that form 2-designs, i.e., maximally expressive ansatze, local costs concentrate exponentially fast to a fixed value. Here, we show that this proof may be generalized to nonperfectly expressive ansatze.

Specifically, we show that for a k-local cost C_k, we have that

$$\langle|C_k - \mathrm{Tr}[(H_k \otimes \mathbb{1})\mathbb{1}/d]|\rangle \leqslant ||H_k||_\infty \left(\sqrt{\int dU_{\mathrm{Haar}} \mathrm{Tr}[(U\otimes U)(\sigma\otimes\sigma)(U^\dagger \otimes U^\dagger)(W\otimes \mathbb{1})]}\right) + ||H_k||_\infty \sqrt{\chi_\epsilon}, \tag{F1}$$

where

$$\left(\sqrt{\int dU_{\mathrm{Haar}} \mathrm{Tr}[(U\otimes U)(\sigma\otimes\sigma)(U^\dagger \otimes U^\dagger)(W\otimes \mathbb{1})]}\right) \in \mathcal{O}\left(\sqrt{\frac{2^k}{2^n}}\right). \tag{F2}$$

Here, χ_ϵ is an expressibility-dependent correction defined as

$$\chi_\epsilon := \mathrm{Tr}[\mathcal{A}_{\mathbb{U}}(|0\rangle\langle 0|)(W\otimes \mathbb{1})], \tag{F3}$$

where, as previously, W is the subsystem permutation operator.

Proof. The start of the proof is identical to Ref. [33]:

$$\begin{aligned}
\langle|C_k - \mathrm{Tr}[(H_k \otimes \mathbb{1})\mathbb{1}/d]|\rangle &= \int dU |((H_k\otimes \mathbb{1})\left(U|0\rangle\langle 0|U^\dagger - \mathbb{1}/d\right))| \\
&\leqslant ||H_k||_\infty \int dU ||\mathrm{Tr}_{\overline{k}}(\left(U|0\rangle\langle 0|U^\dagger - \mathbb{1}/d\right))||_1 \\
&\leqslant ||H_k||_\infty \sqrt{2^k \int dU ||\mathrm{Tr}_{\overline{k}}(\left(U|0\rangle\langle 0|U^\dagger - \mathbb{1}/d\right))||_2^2} \\
&= ||H_k||_\infty \sqrt{\int dU \mathrm{Tr}[\left(U|0\rangle\langle 0|U^\dagger - \mathbb{1}/d\right)\otimes\left(U|0\rangle\langle 0|U^\dagger - \mathbb{1}/d\right)(W\otimes \mathbb{1})]} \\
&= ||H_k||_\infty \sqrt{\int dU \mathrm{Tr}[(U\otimes U)(\sigma\otimes\sigma)(U^\dagger \otimes U^\dagger)(W\otimes \mathbb{1})]},
\end{aligned} \tag{F4}$$

where $\sigma = |0\rangle\langle 0| - \mathbb{1}/d$. The first inequality follows from Hölder's inequality. For the second inequality, we use the relation between the trace norm and the Hilbert-Schmidt norm and invoke Jensen's inequality. We use $\overline{k}$ to denote qubits that are not measured for defining the cost function C_k.

We now substitute in the definition of $\mathcal{A}_{\mathbb{U}}(|0\rangle\langle 0|)$ to obtain that

$$\begin{aligned}
\int dU \mathrm{Tr}[(U\otimes U)(\sigma\otimes\sigma)(U^\dagger \otimes U^\dagger)(W\otimes \mathbb{1})] &= \int dU_{\mathrm{Haar}} \mathrm{Tr}[(U\otimes U)(\sigma\otimes\sigma)(U^\dagger \otimes U^\dagger)(W\otimes \mathbb{1})] \\
&\quad + \mathrm{Tr}[\mathcal{A}_{\mathbb{U}}(|0\rangle\langle 0|)(W\otimes \mathbb{1})].
\end{aligned} \tag{F5}$$

Introducing the shorthand $\mathrm{Tr}[\mathcal{A}_{\mathbb{U}}(|0\rangle\langle 0|)(W\otimes \mathbb{1})] = \chi_\epsilon$ to denote the expressibility-dependent correction, we can then write

$$\begin{aligned}
\langle|C_k - \mathrm{Tr}[(H_k \otimes \mathbb{1})\mathbb{1}/d]|\rangle &\leqslant ||H_k||_\infty \sqrt{\int dU_{\mathrm{Haar}} \mathrm{Tr}[(U\otimes U)(\sigma\otimes\sigma)(U^\dagger \otimes U^\dagger)(W\otimes \mathbb{1})] + \chi_\epsilon} \\
&\leqslant ||H_k||_\infty \left(\sqrt{\int dU_{\mathrm{Haar}} \mathrm{Tr}[(U\otimes U)(\sigma\otimes\sigma)(U^\dagger \otimes U^\dagger)(W\otimes \mathbb{1})]} + \sqrt{\chi_\epsilon}\right),
\end{aligned} \tag{F6}$$

where we use $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$. Moreover, Eq. F2 follows from Theorem 2 in Ref. [70], which completes the proof. ■

APPENDIX G: NUMERICALLY STUDYING THE CORRELATIONS BETWEEN EXPRESSIBILITY AND COST PARTIAL DERIVATIVES

In this appendix, we present numerical results on the correlations between the cost gradient and expressibility. Specifically, we consider the layered parametrized ansatz detailed in Sec. III B of the main text and plot the variance in the PQC gradients as a function of its expressibility.

We can calculate the expressibility measures ε_R^{ρ} and ε_L^{H} via their reformulation in terms of the state- and Hamiltonian-dependent frame potentials $\mathcal{F}_R^{(\rho)} := \mathcal{F}_{\mathbb{U}_R}^{(\rho)}$ and $\mathcal{F}_L^{(H)} := \mathcal{F}_{\mathbb{U}_L}^{(\rho)}$ given in Eq. (B11). However, since it follows from Eq. (B10) that the state-dependent (Hamiltonian-dependent) frame potential for the Haar distribution $\mathcal{F}_{\text{Haar}}^{(\rho)}$ ($\mathcal{F}_{\text{Haar}}^{(H)}$) is exponentially small, and ε_R^{ρ} (ε_L^{H}) measures the difference between $\mathcal{F}_R^{(\rho)}$ and $\mathcal{F}_{\text{Haar}}^{(\rho)}$ ($\mathcal{F}_L^{(H)}$ and $\mathcal{F}_{\text{Haar}}^{(H)}$), it follows that ε_R^{ρ} (ε_L^{H}) may also be exponentially small. We therefore find the ratio of the true frame potential to the Haar frame potential more insightful to plot. That is, we consider the ratios

$$\frac{\mathcal{F}_R^{(\rho)}}{\mathcal{F}_{\text{Haar}}^{(\rho)}} = \frac{(\varepsilon_{U}^{\rho})^2}{\mathcal{F}_{\text{Haar}}^{(\rho)}} + 1, \tag{G1}$$

$$\frac{\mathcal{F}_L^{(H)}}{\mathcal{F}_{\text{Haar}}^{(H)}} = \frac{(\varepsilon_{U}^{H})^2}{\mathcal{F}_{\text{Haar}}^{(H)}} + 1. \tag{G2}$$

The larger these ratios, the more inexpressive is the ansatz, with the ratios tending to 1 for maximally expressive ansatze (exact 2-designs).

In Figs. 7 and 8, we plot the variance in the partial derivative as a function of $\mathcal{F}_R^{(\rho)}/\mathcal{F}_{\text{Haar}}^{(\rho)}$ and $\mathcal{F}_L^{(H)}/\mathcal{F}_{\text{Haar}}^{(H)}$, respectively. In line with Sec. III B of the main text, we focus on three different ways of tuning the expressibility of an ansatz; namely decreasing the depth of the circuits, correlating circuit parameters, and restricting either the direction of rotations or rotation angle ranges.

To numerically quantify the degree of correlations between the variance in the partial derivative of the cost and the expressibility, we include in Figs. 7 and 8 the Spearman correlation coefficient [71] and its corresponding p value. Overall, we find a clear correlation between partial derivatives of the cost and expressibility, with the variance in the derivatives increasing with increasing $\mathcal{F}_R^{(\rho)}$ and $\mathcal{F}_L^{(H)}$. Specifically, combining all the different ways of tuning the expressibility, the Spearman coefficient for the correlation between the variance in the partial derivative of the cost and the ϵ^{ρ} is found to be 0.78, with a p value of 1.19×10^{-7}. Similarly, the Spearman coefficient for the correlations with ϵ^{H} is 0.80, with a p value of 1.18×10^{-7}.

It is noteworthy that the Hamiltonian-dependent frame potential captures the effect of locality on cost gradients as the circuit depth is tuned. As observed in Sec. III B, increasing the depth of the circuit reduces cost partial derivatives for a local cost but not a global cost. The state-dependent frame potential cannot capture this effect, since it is independent of the choice of measurement operator H and therefore necessarily independent of the locality of H. Conversely, while the Hamiltonian frame potential for a local cost decreases with increasing depth, in line with the decreasing variance in partial derivatives, the Hamiltonian-dependent frame potential for the global cost is effectively

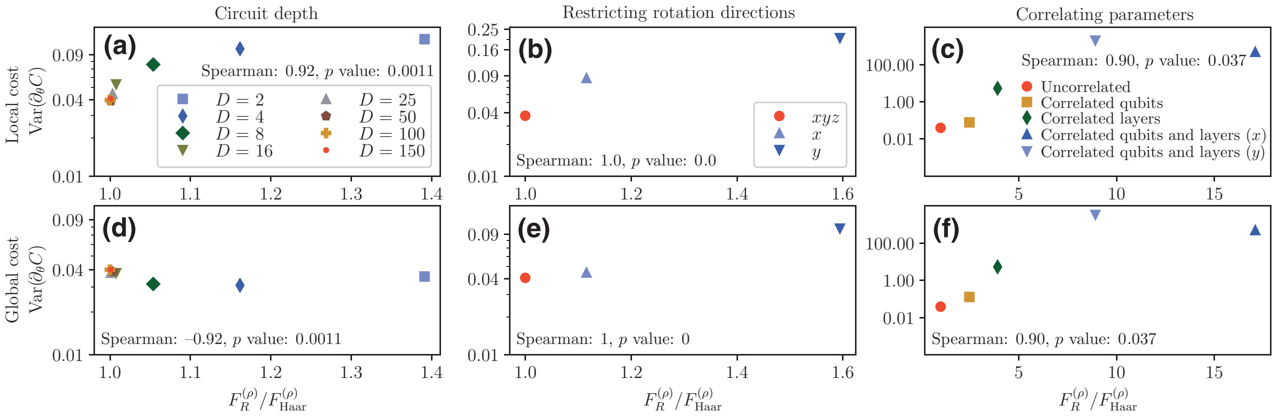

FIG. 7. Correlations between cost partial derivatives and the *state*-dependent frame potential. The variance in the partial derivative of a global cost with $H = \prod_{i=1}^{n} \sigma_i^z$ (top) and the 2-local cost with $H = \sigma_1^z \sigma_2^z$ (bottom) as a function of the expressibility measure $\mathcal{F}_R^{(\rho)}/\mathcal{F}_{\text{Haar}}^{(\rho)}$ {in both cases, $\rho = |\psi_0\rangle\langle\psi_0|^{\otimes n}$ where $|\psi_0\rangle = \exp[-i(\pi/8)\sigma_Y]|0\rangle$}. In the left panel, we vary the circuit depth D of a hardware-efficient ansatz. In the right (middle) panel, we consider the effect of correlating parameters (restricting the directions of rotation) of a hardware-efficient ansatz with $D = 100$, with the choices of correlations (rotations) indicated in the figure legend. In all cases, $n = 4$, the derivative is taken with respect to θ_D^1, and the variance and frame potentials are estimated using an ensemble of 5000 unitaries.

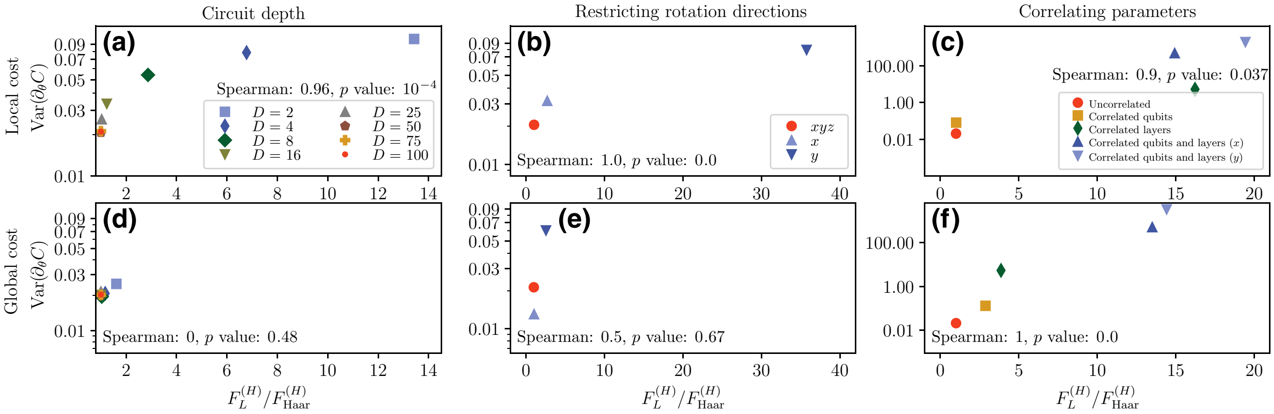

FIG. 8. Correlations between cost partial derivative and the *Hamiltonian*-dependent frame potential. The setting here is entirely equivalent to that described in Fig. 7; however, here we plot the variance in the partial derivative as a function of the ratio of *Hamiltonian*-dependent frame potentials $\mathcal{F}_L^{(H)}/\mathcal{F}_{\mathrm{Haar}}^{(H)}$ and the derivative is taken with respect to θ_1^1.

constant (even as the depth of the circuit substantially increases), reflecting the effectively constant variance in partial derivatives.

Nonetheless, the correlation between the variance in the cost partial derivative and the expressibility is not perfect, as is clear, for example, from Fig. 8(f). This is entirely compatible with our analytic bounds, which are upper bounds and therefore do not enforce perfect correlation between the variance in the partial derivative and the expressibility. Thus while Figs. 7 and 8 demonstrate a clear correlation between the variance in the partial derivative of the cost and expressibility, further work is required to understand the intricacies of this correlation.

[1] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, and *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

[2] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[3] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, eid 4213 (2014).

[4] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, The theory of variational hybrid quantum-classical algorithms, New J. Phys. **18**, 023023 (2016).

[5] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv preprint ArXiv:1411.4028 (2014).

[6] J. Romero, J. P. Olson, and A. Aspuru-Guzik, Quantum autoencoders for efficient compression of quantum data, Quantum Sci. Technol. **2**, 045001 (2017).

[7] S. Khatri, R. LaRose, A. Poremba, L. Cincio, A. T. Sornborger, and P. J. Coles, Quantum-assisted quantum compiling, Quantum **3**, 140 (2019).

[8] R. LaRose, A. Tikku, É. O'Neel-Judy, L. Cincio, and P. J. Coles, Variational quantum state diagonalization, npj Quantum Inf. **5**, 1 (2018).

[9] A. Arrasmith, L. Cincio, A. T. Sornborger, W. H. Zurek, and P. J. Coles, Variational consistent histories as a hybrid algorithm for quantum foundations, Nat. Commun. **10**, 1 (2019).

[10] M. Cerezo, A. Poremba, L. Cincio, and P. J. Coles, Variational quantum fidelity estimation, Quantum **4**, 248 (2020).

[11] K. Sharma, S. Khatri, M. Cerezo, and P. J. Coles, Noise resilience of variational quantum compiling, New J. Phys. **22**, 043006 (2020).

[12] C. Bravo-Prieto, R. LaRose, M. Cerezo, Y. Subasi, L. Cincio, and P. Coles, Variational quantum linear solver, arXiv preprint ArXiv:1909.05820 (2019).

[13] M. Cerezo, K. Sharma, A. Arrasmith, and P. J. Coles, Variational quantum state eigensolver, arXiv preprint ArXiv:2004.01372 (2020).

[14] K. Heya, K. M. Nakanishi, K. Mitarai, and K. Fujii, Subspace variational quantum simulator, arXiv preprint ArXiv:1904.08566 (2019).

[15] C. Cirstoiu, Z. Holmes, J. Iosue, L. Cincio, P. J. Coles, and A. Sornborger, Variational fast forwarding for quantum simulation beyond the coherence time, npj Quantum Inf. **6**, 1 (2020).

[16] B. Commeau, M. Cerezo, Z. Holmes, L. Cincio, P. J. Coles, and A. Sornborger, Variational Hamiltonian diagonalization for dynamical quantum simulation, arXiv preprint ArXiv:2009.02559 (2020).

[17] Y. Li and S. C. Benjamin, Efficient Variational Quantum Simulator Incorporating Active Error Minimization, Phys. Rev. X **7**, 021050 (2017).

[18] S. Endo, J. Sun, Y. Li, S. C. Benjamin, and X. Yuan, Variational Quantum Simulation of General Processes, Phys. Rev. Lett. **125**, 010501 (2020).

[19] X. Yuan, S. Endo, Q. Zhao, Y. Li, and S. C. Benjamin, Theory of variational quantum simulation, Quantum **3**, 191 (2019).

[20] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Variational quantum algorithms, arXiv preprint ArXiv:2012.09265 (2020).

[21] S. Hadfield, Z. Wang, B. O'Gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, Algorithms **12**, 34 (2019).

[22] R. J. Bartlett and M. Musiał, Coupled-cluster theory in quantum chemistry, Rev. Mod. Phys. **79**, 291 (2007).

[23] J. Lee, W. J. Huggins, M. Head-Gordon, and K. B. Whaley, Generalized unitary coupled cluster wave functions for quantum computation, J. Chem. Theory Comput. **15**, 311 (2018).

[24] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, and *et al.*, Quantum chemistry in the age of quantum computing, Chem. Rev. **119**, 10856 (2019).

[25] D. Wecker, M. B. Hastings, and M. Troyer, Progress towards practical quantum variational algorithms, Phys. Rev. A **92**, 042303 (2015).

[26] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature **549**, 242 (2017).

[27] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[28] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, Cost-function-dependent barren plateaus in shallow quantum neural networks, arXiv preprint ArXiv:2001.00550 (2020).

[29] K. Sharma, M. Cerezo, L. Cincio, and P. J. Coles, Trainability of dissipative perceptron-based quantum neural networks, arXiv preprint ArXiv:2005.12458 (2020).

[30] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, arXiv preprint ArXiv:2007.14384 (2020).

[31] M. Cerezo and P. J. Coles, Impact of barren plateaus on the hessian and higher order derivatives, arXiv preprint ArXiv:2008.07454 (2020).

[32] Z. Holmes, A. Arrasmith, B. Yan, P. J. Coles, A. Albrecht, and A. T. Sornborger, Barren plateaus preclude learning scramblers, arXiv preprint ArXiv:2009.14808 (2020).

[33] C. Ortiz Marrero, M. Kieferová, and N. Wiebe, Entanglement induced barren plateaus, arXiv preprint ArXiv:2010.15968 (2020).

[34] A. Uvarov and J. Biamonte, On barren plateaus and cost function locality in variational quantum algorithms, arXiv preprint ArXiv:2011.10530 (2020).

[35] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio, and P. J. Coles, Effect of barren plateaus on gradient-free optimization, arXiv preprint ArXiv:2011.12245 (2020).

[36] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, The power of quantum neural networks, arXiv preprint ArXiv:2011.00027 (2020).

[37] T. Volkoff and P. J. Coles, Large gradients via correlation in random parameterized quantum circuits, Quantum Sci. Technol. **6**, 025008 (2021).

[38] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, An initialization strategy for addressing barren plateaus in parametrized quantum circuits, Quantum **3**, 214 (2019).

[39] G. Verdon, M. Broughton, J. R. McClean, K. J. Sung, R. Babbush, Z. Jiang, H. Neven, and M. Mohseni, Learning to learn with quantum neural networks via classical neural networks, arXiv preprint ArXiv:1907.05415 (2019).

[40] A. Skolik, J. R. McClean, M. Mohseni, P. van der Smagt, and M. Leib, Layerwise learning for quantum neural networks, arXiv preprint ArXiv:2006.14904 (2020).

[41] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger, and P. J. Coles, Absence of barren plateaus in quantum convolutional neural networks, arXiv preprint ArXiv:2011.02966 (2020).

[42] K. Zhang, M.-H. Hsieh, L. Liu, and D. Tao, Toward trainability of quantum neural networks, arXiv preprint ArXiv:2011.06258 (2020).

[43] E. Campos, A. Nasrallah, and J. Biamonte, Abrupt transitions in variational quantum circuit training, arXiv preprint ArXiv:2010.09720 (2020).

[44] T. L. Patti, K. Najafi, X. Gao, and S. F. Yelin, Entanglement devised barren plateau mitigation, arXiv preprint ArXiv:2012.12658 (2020).

[45] K. Bharti and T. Haug, Iterative quantum assisted eigensolver, arXiv preprint ArXiv:2010.05638 (2020).

[46] K. Bharti and T. Haug, Quantum assisted simulator, arXiv preprint ArXiv:2011.06911 (2020).

[47] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms, Adv. Quantum Technol. **2**, 1900070 (2019).

[48] K. Nakaji and N. Yamamoto, Expressibility of the alternating layered ansatz for quantum computation, arXiv preprint ArXiv:2005.12537 (2020).

[49] J. Tangpanitanon, S. Thanasilp, N. Dangniam, M.-A. Lemonde, and D. G. Angelakis, Expressibility and trainability of parameterized analog quantum systems for machine learning applications, arXiv preprint ArXiv:2005.11222 (2020).

[50] In Appendix D, we extend our results to more general costs of the form $C_{\text{gen}} = \sum_i \text{Tr}[H_i U(\boldsymbol{\theta}) \rho_i U(\boldsymbol{\theta})^\dagger]$. This cost allows for multiple input states $\{\rho_i\}$ and measurement operators $\{H_i\}$, opening up quantum machine-learning approaches that employ training data [57–60].

[51] D. P. DiVincenzo, D. W. Leung, and B. M. Terhal, Quantum data hiding, IEEE Trans. Inf. Theory **48**, 580 (2002).

[52] D. Gross, K. Audenaert, and J. Eisert, Evenly distributed unitaries: On the structure of unitary designs, J. Math. Phys. **48**, 052104 (2007).

[53] D. A. Roberts and B. Yoshida, Chaos and complexity by design, J. High Energy Phys. **2017**, 121 (2017).

[54] R. A. Low, *Pseudo-randomness and Learning in Quantum Computation*, Ph.D. thesis, school - (2010).

[55] N. Hunter-Jones, Unitary designs from statistical mechanics in random quantum circuits, arXiv preprint ArXiv:1905.12053 (2019).

[56] A. W. Harrow and R. A. Low, Random quantum circuits are approximate 2-designs, Commun. Math. Phys. **291**, 257 (2009).

[57] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Quantum machine learning, Nature **549**, 195 (2017).

[58] M. Schuld, I. Sinayskiy, and F. Petruccione, An introduction to quantum machine learning, Contemp. Phys. **56**, 172 (2015).

[59] K. Poland, K. Beer, and T. J. Osborne, No free lunch for quantum machine learning, arXiv preprint ArXiv:2003.14103 (2020).

[60] K. Sharma, M. Cerezo, Z. Holmes, L. Cincio, A. Sornborger, and P. J. Coles, Reformulation of the no-free-lunch theorem for entangled data sets, arXiv preprint ArXiv:2007.04900 (2020).

[61] A. Arrasmith, Z. Holmes, M. Cerezo, and P. J. Coles, Equivalence of quantum barren plateaus to cost concentration and narrow gorges, arXiv preprint ArXiv:2104.05868 (2021).

[62] M. Broughton, G. Verdon, T. McCourt, A. J. Martinez, J. Hyeon Yoo, S. V. Isakov, P. Massey, M. Yuezhen Niu, R. Halavati, E. Peters, and *et al.*, TensorFlow Quantum: A software framework for quantum machine learning, arXiv preprint ArXiv:2003.02989 (2020).

[63] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 4812 (2018).

[64] J. L. Myers, A. D. Well, and R. F. Lorch Jr, *Research Design and Statistical Analysis* (Routledge, Oxford, 2013).

[65] We note that for highly correlated circuits, the variance in the gradient can in fact be larger than allowed by our bounds, which are derived for uncorrelated circuits. In future work, it would be interesting to investigate whether our bounds can be generalized to account for such correlations.

[66] W. Brown and O. Fawzi, Scrambling speed of random quantum circuits, arXiv preprint ArXiv:1210.6644 (2012).

[67] A. Harrow and S. Mehraban, Approximate unitary t-designs by short random quantum circuits using nearest-neighbor and long-range gates, arXiv preprint ArXiv:1809.06957 (2018).

[68] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki, Local random quantum circuits are approximate polynomial-designs, Commun. Math. Phys. **346**, 397 (2016).

[69] Z. Puchała and J. Adam Miszczak, Symbolic integration with respect to the Haar measure on the unitary groups, Bull. Pol. Acad. Sci. Tech. Sci. **65**, 21 (2017).

[70] S. Popescu, A. J. Short, and A. Winter, The foundations of statistical mechanics from entanglement: Individual states vs. averages. eprint, arXiv preprint ArXiv:quant-ph/0511225 (2005).

[71] S. C. Choi, Tests of equality of dependent correlation coefficients, Biometrika **64**, 645 (1977).